UNIVERSIDADE CATÓLICA PORTUGUESA

Segmentação dos Postos de Transformação e Distribuição

Trabalho Final na modalidade de Dissertação
apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em gestão com especialização em business analytics

por

Ana Paula Teixeira da Silva

sob orientação de
Prof. Vera Miguéis, Prof. Conceição Portela e Drª Susana Magalhães

Faculdade de Economia e Gestão | Católica Porto Business School
Março 2017

"The truth, as always, will be far stranger"

- Arthur C. Clarke

Agradecimentos

Em primeiro lugar, gostaria de agradecer à Prof. Vera, minha orientadora, toda a disponibilidade que teve ao longo deste percurso, todas as sugestões que me deu e a preocupação constante em acompanhar o meu trabalho na EDP Distribuição. Obrigada também pelos ensinamentos científicos que partilhou comigo ao longo do tempo e pela atitude crítica.

À Prof. Conceição Portela por me ter co-orientado no decorrer desta etapa, pela sua contribuição neste trabalho e ao longo deste meu percurso na Católica Porto Business School.

À Drª Susana Magalhães, minha orientadora na EDP Distribuição, agradeço toda a dedicação e todos os desafios que me foi colocando para que pudesse desenvolver as minhas competências na área do Data Analytics. Agradeço também a companhia e a integração na equipa.

Ao Eng. Mário Lemos, que permitiu a realização deste projeto na Direção de Gestão de Energia na EDP Distribuição, agradeço a oportunidade de realizar este trabalho e a sua colaboração nas minhas tarefas na empresa.

A todos os meus colegas na EDP Distribuição, agradeço todo o companheirismo e apoio proporcionado durante o meu trabalho.

Aos amigos, agradeço o apoio e a motivação. Obrigada pela amizade, companhia e paciência nos tempos mais apertados.

Ao meu Pai, agradeço os conselhos valiosos, a coragem e a motivação para ter seguido este novo rumo na minha vida. Obrigada por acreditares em mim e por todo o apoio que me deste.

Finalmente, à minha Mãe, agradeço toda a motivação e dedicação permanente. Obrigada por te preocupares mais do que qualquer outra pessoa e por poder contar sempre contigo.

Resumo

Este trabalho foi desenvolvido na EDP Distribuição, no âmbito da validação e disponibilização de dados, com o objetivo de colmatar as falhas de registos de potência de consumo nos diagramas de carga dos Postos de Transformação e Distribuição (PTD's) de Energia em Portugal.

Para esse efeito, os PTD's foram segmentados através do algoritmo de *Clustering* K-Means, tendo para isso sido utilizados os valores de potência de consumo registados ao longo do período de um ano. Estes valores de potência foram previamente normalizados para que fosse possível considerar apenas a forma dos perfis aquando da segmentação, e posteriormente agrupados por hora e por dia da semana-tipo em cada mês, com o intuito de reduzir o número de variáveis a processar pelo R.

Da análise de *Clustering*, resultaram três perfis, sob a forma de séries temporais que correspondiam ao período de um ano, representativos de todos os PTD's. Estes perfis foram utilizados para estimar os valores de potência do ano seguinte.

No final, testou-se o procedimento desenvolvido para efetuar as previsões para o mês de Novembro de 2016. Obteve-se um erro entre previsões e valores reais de 27,3%, excluindo possíveis *outliers*, para aquele mês. Considerou-se aceitável este resultado e, portanto, assumiu-se que este método poderia ser utilizado para inferir as falhas de registos de potência de consumo nos diagramas de carga.

Palavras-chave: Postos de Transformação e Distribuição de Energia, Data mining, *Clustering*, Classificação

Abstract

This work was developed at EDP Distribuição, in the field of data validation and availability, aiming to handle failures of electric power consumption records from smart meters data electric power transformer stations (PTD's) in Portugal.

For that purpose, the PTD's were segmented using K-Means (a *Clustering* technique) using the power consumption records recorded over the period of a year. These values were previously normalized, in order to consider only the profile's shape during the segmentation phase, and were grouped by hour and by typical week in each month, in order to reduce the number of variables to be processed by R.

As result of *Clustering*, 3 typical profiles were obtained, under the shape of time series that corresponded to the period of a year, representative of all PTD's. These profiles were used to estimate the power consumption records for the next year.

In the end, the developed procedure was tested to make the predictions for next year's (2016) November. Concerning this month's predictions, it was obtained an error between predictions and real values of 27,3%, excluding possible *outliers*. This result was considered acceptable and thus, it was assumed this procedure could be used to fulfill the defined goals.

Key Words: Smart Meters, Data mining, *Clustering*, Segmentation, Classification

Índice

Índice de Figuras

Índice de Tabelas

Siglas e Abreviaturas

PTD – Posto de Transformação e Distribuição de Energia

EDPD – EDP Distribuição

REN – Redes Energéticas Nacionais

MAT – Muito Alta Tensão

RND – Rede Nacional de Distribuição

MT – Média Tensão

AT – Alta Tensão

BT – Baixa Tensão

TP – Transformador de Potência ou Totalizador

KDD – Knowledge Discovery of Databases

AMR –Automatic Meter Reading

AMI – Automatic Meter Infrastructure

TDLP – Typical Daily Load Profile

SMAS – Smart Meter Data Analytics System

DTW – Dynamic Time Warping

K-NN – K – Nearest Neighbour

DB Index – Davies Bouldin índex

1. Introdução

Este trabalho consistiu na segmentação de Postos de Transformação e Distribuição (PTD's) de energia da rede de distribuição de eletricidade da EDP, com base nos seus diagramas de carga.

A EDP Distribuição (EDPD) está entre os maiores operadores de energia da Península Ibérica. É a única empresa portuguesa que faz parte dos índices Dow Jones de Sustentabilidade, algo que reflete as suas políticas de sustentabilidade e responsabilidade social. Apostou, desde 2015, na instalação de contadores de telecontagem inteligentes ao nível da transformação de energia de média para baixa tensão, algo que contribuiu para que se tornasse uma organização que lida com elevadas quantidades de dados (Big Data). Os registos de potência média (em kW) fornecidos por estes contadores em intervalos de 15 minutos (diagramas de carga) constituíram a base para a realização deste trabalho.

O objetivo desta dissertação reside em agrupar os Postos de Transformação e Distribuição (PTD's) com base nos seus diagramas de carga, utilizando técnicas de *Clustering* e classificação que permitam, respetivamente, criar um grupo de perfis-tipo de consumo e alocar os PTD's restantes a cada um dos *clusters* (agrupamentos) resultantes da segmentação. Deste modo, a EDP Distribuição (EDPD) tem a possibilidade de utilizar os perfis de consumo típicos na validação de dados, para colmatar a inexistência de registos, para detetar valores anómalos ou alterações de padrões de consumo que se traduzam em novas tendências no comportamento dos PTD's.

Durante o estágio foram utilizadas essencialmente duas ferramentas: SQL-Server, devido à sua capacidade de armazenar dados de grandes dimensões e R, devido à forte componente estatística que lhe é inerente e que foi necessária ao desenvolvimento de grande parte deste trabalho.

2. Contextualização no Sector Energético

Em Portugal existem empresas que operam ao nível da produção, transporte, distribuição e comercialização da energia elétrica. A REN exerce atividades de transporte da energia de muito alta tensão (MAT). A EDP Distribuição (EDPD), do Grupo EDP, é concessionária da Rede Nacional de Distribuição (RND) – rede de distribuição de energia elétrica, em Média (MT) e Alta Tensão (AT), no território continental, sendo ainda concessionária, quase na totalidade, da rede de distribuição em Baixa Tensão (BT). A EDPD é então responsável pelo abastecimento de eletricidade, expansão e fiabilidade da rede e fornecimento de serviços aos comercializadores de eletricidade.

Na rede de distribuição, entre os segmentos AT e MT existem Subestações (SE), responsáveis por transformar a energia de alta tensão em média tensão, enquanto entre os segmentos MT e BT existem Postos de Transformação e Distribuição (PTD's), com um ou mais transformadores de potência ou totalizadores (TP). A cada combinação PTD-TP corresponde um diagrama de carga com valores de potência registados em intervalos de 15 minutos.

Os diagramas de carga da EDPD contêm os registos de potência em Quilowatts (kW) fornecidos pelos contadores inteligentes. Os valores de potência a cada 15 minutos permitem construir uma curva representativa, por cada PTD, da variação da energia consumida ao longo do tempo, constituindo uma série temporal. É importante constatar que existem dados anómalos – que contêm registos em falta ou que se desviam do seu padrão regular. Estes são geralmente designados de falhas, causados por problemas técnicos dos contadores, falhas de comunicação, perda de dados, entre outros (Chen, Li, Lau, Cao, & Wang, 2010). Assim, a existência de registos fornecidos pelos contadores que fornecem os diagramas de carga promove a análise de dados em várias áreas das organizações energéticas, nomeadamente no combate à fraude, balanço energético, monitorização e validação de dados e aumento da consciência energética (Kádár, 2011).

No âmbito organizacional, o departamento de garantia de receita utiliza os dados para avaliar potenciais situações de perdas de energia, que podem ser de origem técnica ou não técnica, e a validação de dados possibilita a disponibilização de dados a empresas comercializadoras de energia. Além da utilidade dos contadores para as empresas, os contadores fornecem informação atualizada aos clientes, que podem consultar a utilização de energia, custo e preço em tempo real (Kádár, 2011).

3. Data Analytics

A evolução tecnológica dos últimos anos promoveu a recolha e o armazenamento de grandes volumes de dados, dando origem à era do "Big Data" (Gillon, Brynjolfsson, Mithas, Griffin, & Gupta, 2012) (Mikut & Reischl, 2011).Os estudos mais recentes consideram que Big Data é um termo bastante volátil e que, como tal, pode ser descrito de vários modos(Ylijoki & Porras, 2016). No entanto, a maior parte dos autores concorda que este conceito é definido como um conjunto de dados de elevada dimensão ou complexidade, que pode facilmente ser caracterizado pelos 5 V's - velocidade, veracidade, volume, variedade e variabilidade (Hashem, Yaqoob, Salimah Mokhtar, & Samee Ullah Khan, 2014).Outros autores definem Big Data como uma área de investigação emergente que cobre aspetos relacionados com os dados, mas nunca com a segurança e a privacidade dos mesmos (Ylijoki & Porras, 2016).

É importante referir que em contextos em que são recolhidos grandes volumes de dados, as ferramentas de análise de dados (Data Analytics) ganham extrema relevância, já que possibilitam a realização de análises e previsões que servem de apoio à decisão, melhorando o desempenho das organizações. Valle et al. (2011) referem que organizações de elevado desempenho usam o Data Analytics cinco vezes mais do que organizações de menor desempenho (LaValle, 2011) .

O Data Analytics representa um conjunto de competências, processos e tecnologias que surgiram no sentido de otimizar e inovar os processos das

organizações, para que estas possam atuar rapidamente, suportar a tomada de decisão e reduzir custos, tornando-as assim "*top performers*"(LaValle, 2011). O Analytics representa uma mudança que surge como uma fonte de valor e de vantagem competitiva para as organizações e que permite uma adaptação a novas necessidades e constrangimentos por parte das empresas (Gillon, Brynjolfsson, Mithas, Griffin, & Gupta, 2012) (Acito & Khatri, 2014).

Estudos recentes têm vindo a referir-se ao Data Analytics como uma opção de carreira "*sexy*" (Provost & Fawcett, 2013), já que pode ser utilizado como um conjunto de processos inovadores capazes de interferir na tomada de decisão. Por exemplo, no marketing pode ser usado para diminuição da *churn rate*, diferenciação da publicidade online e cross-selling (Provost & Fawcett, 2013). Apesar de existirem já diversas ferramentas e várias aplicações na área da gestão, é ainda controverso o uso de Data Analytics, nomeadamente pela falta de competências dos recursos humanos das empresas e ainda por alguns custos inerentes ao armazenamento de informação. Assim, as empresas têm vindo a repensar e a reajustar as qualificações dos seus recursos humanos, assim como a orientar a gestão de topo das empresas para o apoio à decisão com base na análise de dados (Holsapple, Lee-Post, & Pakath, 2014).

Alguns estudos, como Gillon (2012), consideram que o Data Analytics poderá estar a criar oportunidades a longo prazo para mudar radicalmente o desempenho das organizações ou, alternativamente, poderá representar apenas um suporte para otimização dos processos das organizações e para a tomada de decisão (Gillon, Brynjolfsson, Mithas, Griffin, & Gupta, 2012).

Compreender o modo como a análise de dados pode ser utilizada para melhorar o desempenho das organizações pode ser um processo complexo e, como tal, torna-se fundamental discriminar três níveis de análise: descritivo, preditivo e prescritivo. Num contexto temporal, a componente descritiva foca-se em reportar o passado, enquanto a preditiva foca-se no passado e no presente para prever o futuro. A componente prescritiva foca-se em apoiar o processo de tomada de decisão, geralmente com recurso a modelos de otimização (Davenport & Patil, 2012).

Data Mining é um ramo da análise de dados que se preocupa com a extração de conhecimento a partir de um grande conjunto de dados (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). É uma etapa que se insere no *Knowledge Discovery in Databases* (KDD), processo que consiste, entre outras coisas, em analisar dados e descobrir padrões. É relevante referir que com a evolução do Data Mining, surgiram várias ferramentas orientadas para a extração de conhecimento e identificação de padrões nos dados, como o RapidMiner, o KNIME e Oracle Data Mining (Mikut & Reischl, 2011).

4. Estado da Arte

A nível mundial, têm vindo a surgir desafios no sector energético que resultam essencialmente do aumento da procura pela energia elétrica e da crescente consciencialização energética ((Ahmad, 2011). Para lidar com este tipo de desafios, as organizações do sector energético começaram a investir numa nova geração de contadores que veio substituir os contadores elétricos convencionais: os contadores inteligentes.

Os contadores inteligentes ou *smart meters* foram introduzidos entre 1980 e 1990 nos Estados Unidos e a sua utilização tem vindo a aumentar rapidamente nos últimos anos. Estima-se que o mercado global dos contadores inteligentes cresça para 20 mil milhões em 2018 (Alejandro, et al., 2014). Prevê-se também que pelo menos 80% da União Europeia venha a instalar contadores inteligentes até 2020 (Flath, Nicolay, Conte, van Dinther, & Filipova-Neumann, 2012). A instalação dos contadores está prevista no âmbito das políticas estabelecidas pelo EU 20-20-20 que incluem uma legislação estabelecida pelo Parlamento Europeu com o objetivo de diminuir em 20% a emissão de gases poluentes para a atmosfera, com vista à consequente diminuição do efeito de estufa, ao aumento da eficiência energética em 20% e ao aumento da utilização de fontes de energia renováveis em 20% ((ESMIG), 2011). Convém ainda referir que os maiores mercados de contadores inteligentes são a América do Norte, Europa e

Ásia Oriental, sobretudo a China. De salientar que as empresas americanas se encontram entre os maiores produtores mundiais de contadores e fornecem o mercado mundial (Alejandro, et al., 2014).

Torna-se então importante contextualizar o aparecimento destes contadores nos países desenvolvidos e nos países em desenvolvimento e, também, na era tecnológica atual, já que a sua instalação e manutenção implicam grandes investimentos e envolvem despesas na ordem dos milhares de milhões de euros. A sua instalação não consta apenas das preocupações das empresas operadoras de energia e dos clientes, mas também das preocupações dos governos relacionadas com a poupança de energia e utilização eficiente e consciente da energia elétrica. No entanto, apesar do estabelecimento de medidas legislativas, como as que estão incluídas no EU 20-20-20, crê-se que, em muitos países, nomeadamente países em desenvolvimento, ainda não existam incentivos por parte dos governos para investir nos contadores, já que é bastante dispendioso investir nas tecnologias de informação e nas infraestruturas de comunicação associadas.

Prevê-se que a procura de energia aumente consideravelmente nos próximos anos, tornando-se imprescindível obter uma rede complexa de contadores que permitam reduzir ao máximo as perdas e responder à procura. Assim, uma das maiores preocupações do sector energético e dos governos é preservar o balanço energético, ou seja, diminuir as diferenças entre a energia fornecida e a energia consumida, com o intuito de manter a eficiência energética. No caso da União Europeia, esta lançou uma diretiva para 2016, em que se requeriam poupanças de energia correspondentes a 9%, na tentativa de tornar a distribuição de energia mais eficiente, dada a escassez dos recursos não renováveis, utilizados como fonte de energia, e dadas as perdas que acompanham os desperdícios de energia. Por vezes, as diferentes questões reguladoras dos países da União Europeia dificultam a aplicação de medidas orientadas para a poupança de energia, mas existem já certas regiões de alguns países que estão a servir de exemplo para impor e testar estas medidas (Flath, Nicolay, Conte, van Dinther, & Filipova-Neumann, 2012). Em oposição ao

contexto europeu descrito, e de acordo com Depuru SS et al., as empresas norte-americanas recebem incentivos para vender eletricidade. Para obterem estes incentivos, estas empresas acabam por não orientar os consumidores para a poupança energética (Depururu, Wang, & Devabhaktuni, 2011).

Além do que já foi referido, existem problemas relacionados com a privacidade e com a segurança, nomeadamente no setor residencial, que têm atrasado ou desmotivado a instalação dos novos contadores de eletricidade. São contestados os direitos que as organizações energéticas têm de aceder e avaliar o comportamento dos consumidores nas suas residências ou até mesmo a própria presença (Depururu, Wang, & Devabhaktuni, 2011).

Apesar dos desafios que a instalação dos contadores apresenta para governos, organizações produtoras, distribuidoras e comercializadoras de energia elétrica e consumidores, as suas funcionalidades têm vindo a apresentar elevada utilidade para as organizações. Antes da existência dos contadores, a informação que estas possuíam, relativamente ao consumo energético, respeitava a dados históricos acumulados pelos distribuidores de energia, utilizados para investigação (De Silva, Yu, Alahakoon, & Holmes, 2011). A partir do momento em que a telecontagem surgiu, e com ela, o fornecimento de dados de forma remota, os contadores sofreram várias modificações nomeadamente devido à evolução da Internet e da tecnologia. Convém referir que apesar do termo *smart meters* ser usualmente utilizado relativamente a contadores de energia elétrica, também pode ser usado na monitorização da água e do gás.

Existem, até ao momento, dois tipos de tecnologia de contadores de energia elétrica: Automatic Meter Reading (AMR) e Automatic Metering Infrastructure (AMI). Enquanto os primeiros utilizam comunicação monodirecional, os segundos utilizam comunicação bidirecional, sendo capazes de transmitir informação útil, quer às organizações quer aos clientes, e permitindo às organizações definir ou alterar certos parâmetros à distância (Alejandro, et al., 2014). Os registos fornecidos pelos contadores têm atingido uma granularidade cada vez maior ao longo do tempo, possibilitando a obtenção da potência

consumida em intervalos de 15 minutos. A partir destes registos podem ser construídos perfis típicos dos contadores, que constituem séries temporais relativas à variação da potência ao longo do tempo (Liu & Nielsen, 2015).

Algo que veio reforçar a utilidade dos contadores foi o aparecimento do Data Analytics e, com isto, a possibilidade de analisar grandes volumes de dados, com uma granularidade tão fina (Liu & Nielsen, 2015), de forma rápida e eficaz. De destacar que os *data scientists* nem sempre lidam com diagramas de carga com intervalos de reduzida granularidade e que nem sempre é fácil lidar com dados em intervalos de 15 minutos, sem agregar previamente a informação fornecida, através de medidas agregadoras, como a média. Na Alemanha, por exemplo, os registos são obtidos em intervalos de 15 minutos ao nível da indústria e de clientes com um consumo anual igual ou superior a 100.000 kW.h. No entanto, para consumidores ao nível residencial e comercial, que apresentam consumos inferiores, o consumo é medido apenas anualmente, demonstrando que nem sempre a granularidade da análise é tão fina (Flath, Nicolay, Conte, van Dinther, & Filipova-Neumann, 2012).

4.1. Utilização dos dados obtidos na segmentação

A utilidade dos contadores reflete-se, entre outros, na previsão do consumo de eletricidade, na extração de perfis de consumo, na segmentação de clientes, na segmentação de postos de distribuição e na transmissão de informação personalizada aos clientes (Flath, Nicolay, Conte, van Dinther, & Filipova-Neumann, 2012) (De Silva, Yu, Alahakoon, & Holmes, 2011) (McLoughlin, Duffy, & Conlon, 2015).

Tendo em consideração os fatores de utilidade referidos, pode-se concluir que existem duas vertentes relativamente à segmentação do mercado energético: a vertente que se foca na segmentação de clientes, por exemplo baseada nos termos contratuais assinados pelo consumidor com a empresa comercializadora, e a vertente que utiliza os diagramas de carga dos postos de transformação e distribuição para segmentar os postos consoante a distribuição

geográfica, climática, entre outros. Por exemplo, Albert, A. et al (2011), definiram distribuições da procura de energia típicas a partir da segmentação de diagramas de carga. Para esse feito, foram agrupadas classes de distribuições da procura energética (PDD's) com distribuições estatisticamente semelhantes, conseguindo, deste modo, obter segmentos ou *clusters* de consumidores que apresentam um comportamento de consumo energético semelhante entre eles. A técnica utilizada para o *Clustering* foi o K-Means, descrito em 4.1.1 da Secção 4. Data Analytics (Albert, Rajagopal, & Sevlian, 2011). Outros autores, como Kim, Young-Il et al (2009), utilizam também o método de *Clustering* K-Means para segmentar os clientes com base em perfis diários de carga típicos. Numa primeira etapa, anterior ao *Clustering*, estes autores segmentam manualmente todos os consumidores que possuem AMR de acordo com atributos como o código de contrato energético entre a empresa fornecedora e o consumidor, gerando a partir desses dados um perfil típico diário (TDLP) para cada grupo. Dentro de cada segmento, é realizado um K-Means, a partir do qual são obtidos grupos intra-segmento sendo os perfis os centróides desses grupos. Estes perfis são aplicados a clientes que não possuem AMR, conforme o segmento (código de contrato energético) a que pertencem, de acordo com certos atributos, como por exemplo o seu consumo médio, permitindo a sua alocação a um dos grupos obtidos. Este trabalho permitiu prever o comportamento de certos segmentos de utilizadores, de modo a apoiar decisões relativamente à potência consumida e instalada (Kim, Shin, Song, & Yang, 2009).

Um dos problemas associados à segmentação está relacionado com a dimensão e granularidade dos dados fornecidos pelos PTD's que, em muitos casos, representam um grande volume. Vários autores têm procurado lidar com este desafio. Em 2015, McLoughlin et al. propuseram um método de *Clustering* cujo objetivo era identificar padrões de consumo de eletricidade semelhantes, para agregar valores de potência e, consequentemente, diminuir o volume de dados (McLoughlin, Duffy, & Conlon, 2015). Os perfis obtidos foram caracterizados através do cálculo da procura de eletricidade média de cada *cluster* num determinado dia. Em 2014, Kwac et al. (2014) propuseram uma

abordagem alternativa que decompõe os padrões de consumo de eletricidade diários em consumo diário total pré-*Clustering* e utiliza um algoritmo K-Means adaptado. Convém referir que, nesta situação, o *Clustering* foi aplicado a mais de 66 milhões de diagramas de carga residenciais, amplificando assim a importância desta técnica de *Data Mining* na análise de grandes quantidades de dados energéticos (Kwac, Flora, & Rajagopal, 2014).

Os métodos de segmentação também podem ser utilizados para obter padrões de consumo energético ao longo do tempo que, podendo estar associados quer a hábitos diários dos consumidores que influenciam o seu consumo energético quer a localizações geográficas (zonas urbanas, industriais, iluminação pública, entre outras), permitem avaliar o possível impacto de fatores socioeconómicos ou geográficos. O objetivo desta segmentação orientada ao consumidor seria criar diferentes programas, consoante o tipo de consumidor, tentando dinamizar a oferta devido a um consumo mais heterogéneo para alguns consumidores ou tentando identificar clientes com consumos excessivos que são energeticamente ineficientes (Albert & Rajagopal, Smart Meter Driven Segmentation: What Your Consumption Says About You, 2013). Por intermédio da segmentação dos consumidores, e da alteração da potência contratada, poder-se-ia modificar o próprio comportamento do cliente no sentido de um consumo energético mais eficiente, reduzindo a potência que lhe estava destinada a nível contratual(Flath, Nicolay, Conte, van Dinther, & Filipova-Neumann, 2012).

Existem, ainda, casos em que os dados recolhidos são utilizados com o intuito de fazer previsões do consumo de eletricidade, no sentido de apoiar a tomada de decisão em áreas como o planeamento e gestão de energia (De Silva, Yu, Alahakoon, & Holmes, 2011).

No âmbito da utilidade dos contadores e dos registos que por eles são fornecidos, surgem várias ferramentas, softwares próprios e técnicas que têm vindo a ser utilizados no setor energético.

Vários softwares, exclusivamente orientados a tratamento de dados do setor energético, têm sido utilizados, essencialmente para análises descritivas do comportamento energético dos consumidores no sector residencial (McLoughlin, Duffy, & Conlon, 2015).

Além das ferramentas orientadas para o sector energético, é popular a utilização de ferramentas que não estão diretamente relacionadas com os contadores de eletricidade mas que se encontram orientadas para a aplicação de técnicas de Data Mining, pela sua versatilidade e diversidade (Mikut & Reischl, 2011). O SPSS e o SAS, orientados para análises estatísticas de dados de qualquer setor, foram os primeiros softwares pagos a ser usados para a análise de dados energéticos (Mikut & Reischl, 2011). Outras ferramentas de acesso livre capazes de lidar com grandes quantidades de dados e de aplicar técnicas de Data Mining, de tratamento e processamento de dados, têm sido usadas, nomeadamente o R e o RapidMiner. Estas têm à sua disposição uma grande variedade de algoritmos, possibilitando a escolha do método mais adequado para o problema em questão (Jovic, Brkic, & Bogunovic, 2014). A oferta destas ferramentas é alargada, principalmente no que toca a técnicas de C*lustering* e Classificação. Enquanto o RapidMiner apresenta essencialmente técnicas gerais de Data Mining integradas num ambiente de trabalho *user friendly*, o R apresenta uma forte componente estatística, e, em geral, implica o uso de uma linguagem própria de programação(Jovic, Brkic, & Bogunovic, 2014). Outras ferramentas que têm vindo a ser referidas no âmbito da aplicação de Data Mining são o (1) Weka, cujos componentes têm vindo a ser integrados noutras ferramentas, como o RapidMiner, desenvolvidos na linguagem de programação Java, (2) Orange e (3) scikit-learn, desenvolvidos na linguagem de programação Python, e (4) KNIME, que permite visualizar uma série de "blocos de construção" à semelhança do RapidMiner e integrar o R e o Weka (Jovic, Brkic, & Bogunovic, 2014)(Mikut & Reischl, 2011).

Mais recentemente, têm surgido plataformas que permitem utilizar ferramentas de Data Mining para lidar especificamente com dados de consumo. Um exemplo disso é o Smart meter data analytics system (SMAS), proposto por

Xiufeng et al. (2015), que é composto por 3 partes diferenciadas: a "ingestão" de dados, o seu processamento e análise (Liu & Nielsen, 2015).

A aplicação de algumas das ferramentas acima referidas tem sido essencial, nomeadamente na realização de segmentação de consumidores ou PTD's através de técnicas de *Clustering*. Estas técnicas serão descritas na Secção seguinte.

4.1.1. Clustering

O *Clustering* é uma técnica de Data Mining que permite agrupar objetos. O *Clustering* é uma técnica não supervisionada, ou seja, baseada somente na informação disponível nos dados sem qualquer indicação do utilizador/investigador relativamente a *outputs* pretendidos, e que consiste no agrupamento dos dados em *clusters* de forma a maximizar as semelhanças intra-*cluster* e as diferenças inter-*cluster* (Halkidi, Batistakis, & Vazirgiannis, 2001).

Existem vários desafios ao utilizar esta técnica (Jain & Verma, 2014), nomeadamente:

- Big Data (o grande volume de dados dificulta o processamento dos dados em certos softwares)
- "Curse of dimensionality" (número elevado de variáveis dificulta o processamento dos dados e restringe as técnicas ou os softwares que poderão ser utilizados)
- Tratamento de *outliers* (nem todos os algoritmos de *Clustering* estão preparados para lidar com *outliers*, e, portanto, é algo a considerar na seleção do algoritmo)
- Ocupação do espaço disponível em memória RAM, utilizado por algumas ferramentas (ex: R)
- Instabilidade técnica, que se deve ao facto de existir a possibilidade de obter *clusters* diferentes a cada iteração, isto é, cada vez que é realizado o *Clustering* (Zerhari, Lahcen, & Mouline, 2015).

- Dificuldade em lidar com *missing values* (os valores em falta são inferidos ou são criadas alternativas para encarar este tipo de valores)

Os métodos de *Clustering* podem ser classificados como: métodos particionais, hierárquicos e de densidade. Em qualquer um destes métodos, é fundamental a escolha do número de *clusters* k. Os métodos particionais dividem as observações em agrupamentos ou clusters não sobrepostos, de acordo com as suas características. Os métodos aglomerativos são úteis em várias circunstâncias porque representam a segmentação sob a forma de uma árvore hierárquica, permitindo assim uma visualização clara do agrupamento dos dados feito(Zerhari, Lahcen, & Mouline, 2015). Nos métodos baseados em densidade, destacam-se o DBSCAN, que não necessita da escolha do número de *clusters* e é capaz de identificar *outliers* (Xiong, 2007).

Existem vantagens em cada um dos métodos de C*lustering* mencionados, mas é importante avaliar as suas aplicações no ramo energético. O K-Means é um método de *Clustering* muito utilizado pelos especialistas. Este método baseia-se no cálculo da distância entre as observações e o centróide dos potenciais *clusters* (Forgy, 1965). O K-Means inicia-se alocando observações a centróides iniciais mais ou menos arbitrários, que não constituem observações. Quando todas as observações estão alocadas ao centróide que se encontra a menor distância (geralmente a distância euclidiana), a posição dos centróides é recalculada e as observações voltam a ser alocadas aos centróides que se encontrem a menor distância, até que os centróides estabilizem. O K-Medoids, utiliza observações existentes como centróides. Neste caso, na primeira iteração, é criada uma matriz das distâncias entre todas as observações, incluindo os centróides. Para alocação das observações aos centróides, o algoritmo vai buscar as distâncias que necessita à matriz já existente. Este método é eficiente no que respeita ao tempo de execução visto que não há necessidade de recalcular as distâncias a cada iteração. No entanto, ocupa muita memória devido à alocação inicial e permanente da matriz das distâncias, algo que pode constituir um problema com dados de grande dimensão (Kaufman, 1987). Nem

sempre algoritmos como o K-Means e o K-Medoids utilizam a distância euclidiana para avaliar a proximidade das observações. No caso das séries temporais, é muito comum a utilização do Dynamic Time Warping (DTW) para o cálculo das distâncias. O DTW é uma técnica que permite comparar duas séries temporais, que podem ou não variar em velocidade. Para melhor o compreender, assumam-se, por exemplo, duas séries temporais: uma com x1, x2, x3 e x4 instantes temporais e outra com y1, y2, y3 e y4 instantes temporais. Considerando a distância euclideana, os algoritmos usam a distância entre x1 e y1, x2 e y2, x3 e y3 e x4 e y4. Com o DTW, os algoritmos usam a menor distância entre as duas séries. Ou seja, a distância é estimada usando os instantes que estiverem a uma distância menor, sendo que por exemplo, y2 e y3 podem estar ligados a x2 simultaneamente, ainda que y3 esteja desalinhado com x2, desde que a distância seja a mais curta (Giorgino, 2009) (ver Figura 1 para uma melhor compreensão). Assim, no âmbito do *Clustering*, podem ficar no mesmo cluster séries temporais com instantes iniciais e finais diferentes, ou mesmo com diferente número de instantes temporais, desde que, por exemplo, possuam a mesma forma. É possível, nalguns casos, especificar o alcance ou janela de pesquisa pela maior proximidade das coordenadas. Deste modo, assegura-se que o desalinhamento não seja demasiado grande entre observações.

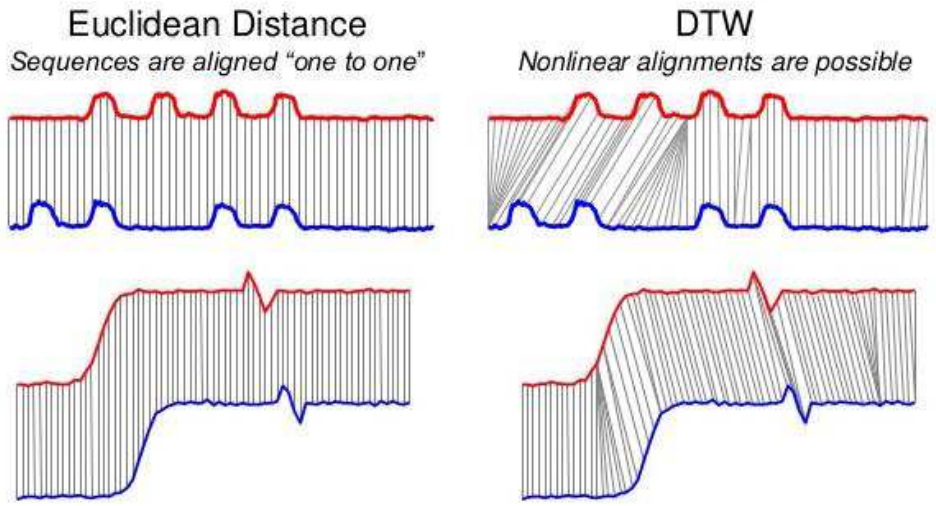


Figura 1

Esquema comparativo que permite ilustrar a estimativa da distância euclidiana e da *Dynamic Time Warping*. (Adaptado de http://www.cs.ucr.edu/~eamonn/tutorials.html)

Convém ainda destacar o DBSCAN, método de *Clustering* baseado em densidades, capaz de detetar *outliers.* Embora apresente bom desempenho, requer bastante espaço de memória RAM, algo que tem que ser tido em consideração em situações em que se lida com uma grande quantidade de informação, como é o caso dos diagramas de carga.

É importante mencionar também que existe uma implementação do K-Means que suporta grandes volumes de dados (Big Data) e que vale a pena mencionar. Pode ser instalado no R através da utilização de dois pacotes do R, biganalytics e bigmemory, sendo que esta combinação permite ultrapassar um dos maiores problemas que existe na utilização do R: a utilização da memória RAM.

Neste trabalho, não foram utilizadas observações com *missing values* para a realização do *Clustering.* Assim sendo, após a análise de C*lustering*, recorreu-se a técnicas de Classificação para alocar aos *clusters* obtidos as observações que não foram consideradas no *Clustering*.

As Secções seguintes apresentam uma breve descrição de técnicas de Classificação existentes na literatura.

4.1.2. Classificação

A Classificação é uma técnica supervisionada de Data Mining, que surge frequentemente no seguimento do *Clustering*, para alocar novas observações aos *clusters* obtidos (Verma, Srivastava, Chack, Kumar Diswar, & Gupta, 2012). À semelhança do que acontece com o *Clustering*, surgem problemas na utilização de algoritmos de Classificação devido ao elevado volume de dados utilizados e devido às limitações de processamento de certas ferramentas de Data Mining como o R.

A maior parte das técnicas de Classificação requer a construção de um modelo, baseado nos dados que já existem, que irá ser utilizado posteriormente para realizar previsões. Este modelo tem como *input* um conjunto de dados de treino, isto é, um conjunto de dados que vão ser usados para construir o modelo. Antes de utilizar o modelo para realizar previsões, este é testado

comparando as previsões com dados reais existentes, permitindo avaliar a precisão com que o modelo prevê a classe a que pertence cada observação. Para esse efeito, os dados são divididos, sendo que a maior parte é usualmente utilizada para criar o modelo e os restantes para o testar. Depois de construído e validado, o modelo é utilizado para fazer previsões.

No contexto deste trabalho, torna-se importante destacar a variedade de técnicas que têm vindo a surgir no âmbito da classificação de observações. As técnicas de Classificação podem ser divididas em várias categorias, entre as quais (Kuhn & Johnson, 2013) (Wu, et al., 2007).

- algoritmos lineares que utilizam a regressão
- algoritmos que se baseiam em árvores de decisão (RandomForests e C4.5)
- algoritmos não lineares, como o K-Nearest Neighbors (K-NN) e redes neuronais.

Visto que este trabalho se focou essencialmente na realização do *Clustering* e que a Classificação surgiu apenas como forma de contornar certos desafios relacionados com a existência de *missing values* nos dados fornecidos, segue-se uma breve descrição de dois algoritmos: RandomForest, que, como o nome indica, consiste num conjunto de árvores – Floresta (Forest), e o K-NN.

Referiu-se anteriormente que o modelo era criado apresentando-se-lhe um conjunto de dados de treino. Esses dados são constituídos por várias séries temporais, a que se chamará vetores, e por um conjunto de classificações. A cada vetor corresponde uma só classificação.

No caso do RandomForest, fala-se de árvores de decisão. Uma árvore tem vários nós. Cada nó tem uma dada constante atribuída pelo algoritmo criador do modelo. Para saber qual a classificação de um dado vetor, compara-se a constante do primeiro nó com uma coordenada do vetor, também escolhida pelo algoritmo. Conforme for maior ou menor, vai para um outro nó ou folha, de dois possíveis. Se se tratar de uma folha, esta contém a classificação que lhe corresponde e o processo para. Se se tratar de um novo nó, uma nova

coordenada é comparada com a sua constante, repetindo-se o processo até se encontrar uma folha. Cada uma destas árvores é construída a partir de uma amostra aleatória dos vetores fornecidos para treino. Daí o nome Random. Constroem-se então um grande número de árvores (Forest) desta forma, cada uma de uma amostra diferente, que passam a constituir o nosso modelo.

Finalmente, para classificar um dado vetor, submete-se o mesmo à classificação de cada árvore, conforme descrito acima. Um critério muito usado para selecionar a classificação do vetor é o do maior número de votos. Ou seja, ganha a classificação produzida por mais árvores. No caso de se pretender prever valores também se usa como critério, por exemplo fazer as médias das classificações.

Por sua vez, o K-NN é uma técnica de Classificação baseada na distância a que se encontram os *k* vizinhos mais próximos (nearest-neighbors) de uma observação a classificar. Este algoritmo necessita, tal como o K-Means, que lhe seja providenciado o parâmetro *k*, que neste caso, constitui o número de vizinhos mais próximos de uma determinada observação que se pretende considerar para a classificar. A classe desta observação é dada pela classe da maioria dos vizinhos próximos (Guo, 2003).

5. Objetivo

O objetivo deste trabalho consistiu em agrupar cerca de 66000 Postos de Transformação e Distribuição da EDP Distribuição, com base nos seus diagramas de carga, para:

- Construir perfis típicos para cada segmento obtido
- Validar dados e colmatar falhas de registos.

6. Metodologia

A metodologia aqui descrita tem como objetivo obter perfis representativos dos PTD's que possam ser utilizados para colmatar as falhas de registos de potência existentes nos seus diagramas de carga. Tendo isso em consideração, os procedimentos que se seguem focaram-se em segmentar os PTD's com base nos seus diagramas de carga, utilizando técnicas de *Clustering*. No caso de existirem PTD's com registos em falta nos seus diagramas de carga, a metodologia adotada consistiu em identificar, por intermédio de técnicas de Classificação, qual o *cluster* a que estes pertenciam.

Ao longo das etapas seguintes, foi necessário lidar com a ocorrência de falhas na contagem (*missing values*), fator que obrigou a que o tratamento de dados fosse necessariamente mais complexo e cuidado. Nem todos os PTD's com falhas foram excluídos do *Clustering*. Com o intuito de obter uma amostra o mais representativa possível para a segmentação e com o maior número de PTD's possível, foram estimados valores de potência em falta, referentes a períodos curtos (até 1h).

Outro fator de grande influência na metodologia utilizada foi a dimensão da base de dados usada, que restringiu ou dificultou a realização de alguns procedimentos no R e no SQL Server. Tendo em conta esta dificuldade, houve a necessidade de reduzir o número de variáveis a considerar para o *Clustering*.

Torna-se importante referir também que nem todos os PTD's alvos de análise tiveram telecontagem instalada no mesmo momento, pelo que, para realização do *Clustering*, foram considerados apenas os PTD's com registos no período considerado, correspondente a um ano entre Novembro de 2015 e Outubro de 2016.

A metodologia descrita divide-se nas seguintes etapas:

- Definição de métodos para o cálculo de estimativa de valores de potência em falta, execução de testes para diferentes durações de falha e medição dos respetivos erros;

- Implementação do método de estimativa com menor erro observado e para a duração máxima permitida (decisão da área de negócio, em função dos resultados dos testes);
- Preenchimento de valores em falta;
- Normalização dos dados da potência;
- Testes de agrupamento, com o intuito de reduzir variáveis;
- Agregação dos dados em função dos testes anteriores;
- *Clustering* dos PTD's que se encontram no mesmo horizonte temporal e que não possuem valores em falta após o preenchimento acima referido;
- Classificação ou alocação dos PTD's não considerados para o *Clustering* nos *clusters* obtidos na etapa anterior;
- Cálculo de estimativas de potência para o ano seguinte;
- Cálculo do erro entre as potências reais obtidas e as estimativas obtidas para um mês, com o intuito de avaliar a viabilidade da metodologia desenvolvida.

Os dados utilizados são estruturados e encontram-se numa base de dados relacional no SQL Server. A tabela inicial de dados continha informação sobre o código dos PTD's, o TP ou Totalizador, a data e a potência registada, entre outras. Para melhor compreensão dos procedimentos que se seguem convém referir que as datas podem ser designadas de variáveis temporais quando consideramos os vários registos de potência ao longo de um período de tempo (série temporal).

6.1. Estimativas de potência

Nesta etapa, pretende-se determinar um método de preenchimento (interpolação) de valores em falta, que verifique determinadas condições de duração e momento de ocorrência das falhas de registo.

Para tal, procedeu-se a várias experiências usando uma amostra aleatória de PTD's com dados completos. Removeram-se fragmentos aleatórios desses dados, de modo a simular falhas nas condições acima referidas. Posteriormente aplicou-se um dos métodos a selecionar para o preenchimento e avaliou-se, por comparação com os dados originais, o seu grau de adequação.

Para esse efeito, tiveram-se em conta vários elementos:

- tamanho da amostra utilizada para simular a inexistência de registos;
- escolha de dois métodos utilizados para calcular essas mesmas estimativas (*média* e *regressão linear*);
- duração das falhas a estimar (15, 30, 45 e 60 minutos);
- número de registos ou pontos de estimativa utilizados para estimar os registos em falta;
- período do dia em que ocorrem as falhas, já que, em trabalhos anteriores realizados pela EDP Distribuição, se tem verificado a ocorrência de variações de consumo ao longo do dia.

Foi utilizada uma amostra de apenas 1000 PTD's com dados referentes ao mês de Julho de 2016, amostra esta que se considerou apresentar um número suficiente de registos para efetuar os testes pretendidos. Como já se referiu, só foram considerados PTD's que possuíam todas as leituras (sem falhas) ou seja com os 96 registos de 15 minutos diários.

Foi também tido em conta que poderiam existir variações, no cálculo dos valores de potência em falta, dependentes do período do dia. Assim, o dia foi dividido em 6 períodos, com o intuito de averiguar se os erros das estimativas variavam muito quer a nível global, quer por período do dia. Os períodos considerados foram:

- Manhã I, das 06h00 às 10h00;
- Manhã II, das 10h00 às 12h00;
- Meio do Dia, das 12h00 às 14h00;

- Tarde, das 14h00 às 20h00;
- Resto do Dia, das 20h00 às 00h00;
- Noite, das 00h00 às 06h00.

Nesse sentido, foi gerada uma amostra de N números aleatórios que permitisse selecionar 10, 100, 1000 ou 10000 registos de PTD's por período, ou seja 60, 600, 6000 ou 60000 registos na totalidade, do conjunto de dados completo. Estes registos foram retirados à amostra como se de falhas se tratassem (simulação de falhas), podendo assim ser estimados a partir dos métodos em avaliação.

Tendo-se o valor real e a estimativa, quer o erro absoluto percentual global quer o erro absoluto percentual em cada período foram calculados através da fórmula:

$$\frac{|estimativa - valor\ real|}{valor\ real} \text{ x } 100\%$$

Os métodos matemáticos a avaliar para o cálculo das estimativas foram a Média e a Regressão Linear. No caso da Média, foram utilizados os valores de potência imediatamente anteriores e posteriores ao instante do registo em falta. No caso da Regressão Linear, para o cálculo do erro percentual absoluto, foi necessário calcular o declive ($\boldsymbol{m}$) e a ordenada na origem ($\boldsymbol{b}$) da reta, utilizando as fórmulas matemáticas habituais para calcular estes parâmetros. Assim, para os valores de x em falta (instantes temporais) tornou-se possível calcular os respetivos valores estimados de y (estimativas da potência).

Consideraram-se intervalos com valores em falta de 15, 30, 45 e 60 minutos. O processo de simulação destas falhas consiste em gerar um momento de falha correspondente a um dado período de 15 minutos, de forma aleatória, mas obedecendo aos constrangimentos já referidos. No caso da duração da falha simulada ser de 15 minutos, o valor a considerar é o registo correspondente a esse período de 15 minutos. Para falhas simuladas de 30, 45 e 60 minutos, os

valores a considerar incluem o valor correspondente àquele período e os correspondentes a 1, 2 ou 3 períodos anteriores respetivamente.

Estabelecido o modo como se geravam as falhas hipotéticas, foi necessário estabelecer quantos períodos de valores existentes, podiam ser usados para estimar a potência. Considerou-se a hipótese de utilizar 1, 2 e 3 registos desses valores, ou pontos para estimativa, correspondentes aos períodos existentes, imediatamente anteriores e posteriores à zona de falha. A ideia seria averiguar qual o número desses pontos para estimativa mais indicado para estimar os valores de potência em cada método e para cada duração de falha, ou seja, o número de pontos que correspondesse a um erro menor no cálculo da estimativa.

É importante referir que se eliminaram os casos em que a potência medida era nula, pois esta aparece como denominador na fórmula do erro percentual absoluto utilizada, tornando impossível o cálculo deste erro. Além disso, também foram removidos os casos em que os períodos correspondentes às falhas geradas ocorreram nos extremos da amostra (ordenada pela Data) uma vez que não é exequível a interpolação nestas zonas por nenhum dos dois métodos que considerámos. A contagem de registos final foi inferior aos valores 60, 600, 6000 e 60000 inicialmente estabelecidos. De referir, ainda, que os pontos utilizados para o cálculo da estimativa por período podiam encontrar-se no período anterior ou no período seguinte, desde que a duração da falha (os registos retirados aleatoriamente) se encontrasse somente no período em causa.

Todo o processo de cálculo dos erros percentuais absolutos globais e por período foi efetuado no SQL Server. A sua análise, na secção *Resultados e Discussão*, irá permitir a recomendação do(s) método(s) para estimar os valores de potência em falta nos diagramas de carga dos PTD's, do número de pontos para estimativa a utilizar para cada duração de falhas e da escala a que devem ser realizadas as estimativas (global ou por período).

6.2. Preenchimento de *Missing values* e Normalização Dados

Numa segunda fase deste trabalho, na tentativa de reduzir os *missing values* (ou falhas) presentes na base de dados, procurou-se estimar os valores de potência associados a falhas mais curtas (no máximo uma hora).

O cálculo das estimativas dos valores de potência de falhas não superiores a 1h foi realizado no SQL Server. Para esse efeito, foram identificadas as falhas existentes e incrementados registos, na tabela de dados, na condição de que as falhas eram inferiores ou iguais a 1h. Para a estimativa, foi utilizado o método de regressão linear que recorreu a 2 pontos de estimativa, dois de cada lado da falha.

Depois de estimadas as falhas inferiores ou iguais a 1h, foi realizada a normalização dos dados referentes às potências de cada PTD, a fim de eliminar o efeito da sua magnitude nos procedimentos seguintes, quer ao nível da análise exploratória quer ao nível do *Clustering*.

O método de normalização utilizado, executado no SQL Server, foi o de converter a série temporal de cada PTD no intervalo [0,1], tendo-se usado para o efeito a seguinte fórmula:

$$potência\ normalizada = \frac{potência - potência\ mínima}{potência\ máxima - potência\ mínima}$$

6.3. Análise Exploratória de Dados e Testes de Agrupamento

A análise exploratória dos dados (AED) foi realizada com dois objetivos:

- Provar que é possível agrupar manualmente PTD's e que, após segmentação, estes apresentam perfis típicos diferentes.

- Reduzir variáveis no horizonte temporal, etapa fundamental para o *Clustering* devido ao grande volume de dados envolvido.

I) Sendo um dos objetivos agrupar manualmente os PTD's, para analisar os perfis resultantes, foi estabelecida, em cada análise, uma margem de 20% entre dois subgrupos de PTD's, ou seja, se o objetivo fosse avaliar o perfil dos PTD's ao longo de 24 horas, comparavam-se PTD's que apresentassem um consumo 20% superior de dia relativamente à noite, com aqueles que apresentassem um consumo 20% superior durante a noite.

A AED foi realizada com os dados relativos ao mês de Julho (mês escolhido aleatoriamente) e a cerca de 1000 PTD's. Para esse efeito, foram utilizados o Microsoft Office Excel, o SQL Server e o R. Esta análise dividiu-se em várias partes, tendo consistido em comparar Dia *versus* Noite, Dias Úteis *versus* fins de semanas ao longo da semana e ao longo do mês.

II) A fase seguinte da AED consistiu em reduzir variáveis no horizonte temporal, devido à elevada dimensionalidade dos dados (número de variáveis temporais a considerar para o *Clustering* que seria de 96 registos diários x 365 dias por ano, correspondendo a um total de 35040 variáveis temporais.

Assim, a segunda etapa da AED consistiu em realizar testes estatísticos em R que nos permitissem avaliar se é possível agrupar os dados (de granularidade inicial de 15 minutos) segundo as horas, segundo os dias e/ou segundo os dias da semana-tipo em cada mês, usando a *média* como função agregadora.

Para esta etapa, por questões de espaço na memória RAM do servidor utilizado, a realização dos testes de agrupamento foi realizada utilizando um mês de cada vez.

Para a aferição da relação entre os valores de carga entre instantes temporais diferentes, utilizou-se o cálculo do cosseno entre dois vetores referentes a esses valores de carga, v e vt, para avaliar a similaridade entre os dois. Quanto mais próximo de 1 estiver o valor do cosseno, maior similaridade existe entre os dois

vetores numa perspetiva multidimensional. O vetor *v* continha os registos de 15 em 15 minutos, e o vetor *vt* continha as médias dos dados agrupados pelo intervalo de tempo determinado em cada etapa: hora, dia, dia da semana-tipo em cada mês.

A Tabela 1 exemplifica os dados que continham os dois vetores, se se considerassem 8 registos agrupados por hora, permitindo uma melhor compreensão do que foi feito. Considera-se que uma só potência média passa a ser representativa de cada hora, ou seja, de cada 4 períodos de 15 minutos.

Tabela 1

Esquema do Agrupamento por Hora dos registos de potência, inicialmente com intervalos de 15 em 15 minutos.

Registos de 15 em 15 minutos (v)	**Registos agrupados por hora (vt**
R1	Média 1
R2	Média 1
R3	Média 1
R4	Média 1
R5	Média 2
R6	Média 2
R7	Média 2
R8	Média 2

Dado que os testes validaram o agrupamento, os dados foram agrupados por hora e por dia da semana-tipo em cada mês. A função de agrupamento foi a *média*.

6.4. *Clustering*

Após importação dos dados para o R, estes foram tratados para que se tornasse possível a realização do *Clustering*. Para esse efeito, foi redefinida a identificação de cada combinação CodigoPTD e Totalizador como PTD_ID e foi modificada a forma da tabela, no sentido de ter um PTD_ID por linha e um intervalo temporal por coluna (variáveis temporais), como exemplificado na Figura 2. Convém referir que os valores que se encontram em Δt_1, Δt_2 até Δt_n na figura, correspondem a uma série temporal constituída pelos valores de potência normalizados.

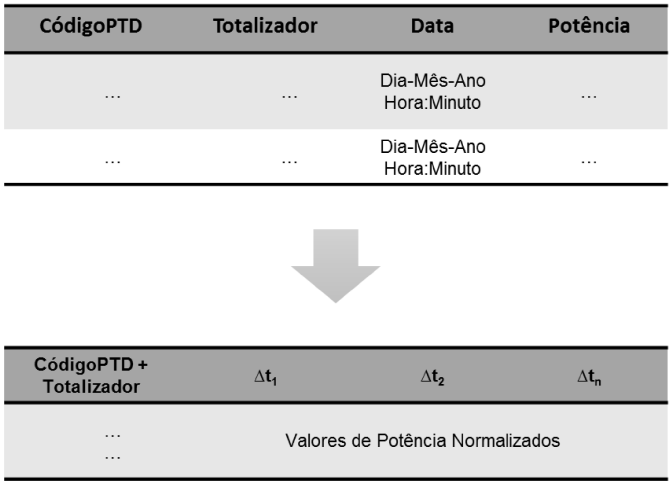


Figura 2

Esquema da transformação *dcast* do data.frame dos diagramas de carga presentes no SQL Server no data.frame modificado para a realização do *Clustering*.

Os dados considerados para o *Clustering* correspondiam a 58058 PTD's com dados completos agrupados por hora e por dia da semana-tipo em cada mês entre Novembro de 2015 e Outubro de 2016. Isto significa que cerca de 85% dos PTD's foram considerados para o *Clustering* e que 10442 PTD's (cerca de 15%) seriam posteriormente classificados nos *clusters* obtidos.

O *Clustering* foi realizado utilizando o método particional K-Means no R. Uma vez que este método implica a definição do número de *clusters* a considerar, recorreu-se a dois métodos para a determinação do número ótimo de *clusters:* o Elbow Method e o Davies Bouldin index (DB index). Os dois

métodos revelaram que o número adequado de clusters a considerar deveria de ser 3. A medida de distância considerada, na execução do algoritmo, para avaliar a semelhança entre observações foi a distância euclideana.

Seguidamente, os resultados obtidos em R foram exportados para o Microsoft Office Excel, possibilitando uma análise dos perfis obtidos.

Para terminar esta fase, pretendeu-se verificar se a segmentação realizada globalmente para o período de um ano (de Novembro de 2015 a Outubro de 2016) se mantinha quando efetuada no período de um mês. Para esse efeito procedeu-se à escolha de dois meses com potencialmente consumos energéticos heterogéneos e distantes no tempo – Setembro e Abril. Realizou-se a segmentação (*Clustering*) para cada um dos dois meses, tendo-se determinado quantos estavam em *clusters* diferentes relativamente à segmentação anual. Na Secção *Resultados e Discussão*, é revelado que, de facto, a segmentação ao longo do ano é representativa da dos meses de Setembro e Abril. Tendo isto em consideração, os *clusters* obtidos para o período de Novembro de 2015 a Outubro de 2016 foram considerados para análise e para a fase seguinte deste trabalho – a Classificação.

6.5. Classificação

Nesta etapa do trabalho, o objetivo foi identificar, para os PTD's que possuíam *missing values* e não foram incluídos no *Clustering* (10442 PTD's), qual o *cluster* a atribuir a cada um deles.

Tendo em conta que estes PTD's apresentaram *missing values,* correspondentes falhas de registo em diferentes instantes do ano e com durações variadas, não foi considerado todo o período de Novembro de 2015 a Outubro de 2016 para a Classificação. Para cada PTD foram apenas considerados os instantes temporais em que havia registos de carga e, como consequência, cada um dos três *clusters* foi também caracterizado apenas com base nos referidos instantes temporais. Assim, ocorrendo os *missing values* em instantes diferentes para cada PTD, definiu-se um modelo de classificação

diferente para cada um deles, pois não podia ser gerado um modelo global, aplicável a todos os PTD's.

A Classificação foi realizada em R, utilizando o algoritmo *Random Forest* já descrito. Foram definidas 100 árvores de decisão (ntree=100), sendo que com 1000, o processo era extremamente lento devido à quantidade de dados envolvidos. A utilização do Random Forest implicou ainda a definição do número de variáveis a considerar em cada *split* dos nós da árvore de decisão. Esse valor foi definido como 6. Relativamente aos outros parâmetros, assumimos as definições por defeito do módulo do R.

Adicionalmente, foi utilizado um método mais simples de tentar identificar qual o *cluster* a que pertence cada PTD. Este método consistiu em determinar qual o centróide mais próximo do vetor correspondente ao PTD (menor distância euclidiana). O cluster desse centróide foi então atribuído a esse PTD. Este processo tem a vantagem de ser mais rápido pois apenas tem de calcular a distância euclidiana entre todos os centróides e os vetores correspondentes a cada um dos 10442 PTD's.

Como resultado desta etapa e do *Clustering*, obteve-se a distribuição dos PTD's (completos e incompletos) pelos 3 *clusters*. Assim, assumiu-se que as séries temporais dos PTD's seriam representadas pelas séries temporais ou perfil dos centróides representativos de cada *cluster.* Visto que estes perfis encaram apenas a forma e não a escala, já que os valores de potência foram normalizados, a última etapa do meu trabalho consistiu em desenvolver uma metodologia que possibilitasse a construção de uma série temporal ou perfil típico com valores de potência à escala para o ano seguinte (Novembro de 2016 a Outubro de 2017), com base nos valores dos centróides obtidos. Estas previsões ou estimativas poderiam ser utilizadas posteriormente para colmatar as falhas de registo.

6.6. Construção dos perfis com base nos resultados do *Clustering*

O objetivo desta etapa consistiu em construir perfis dos *clusters* (os seus centróides) para se obterem os perfis de potências por PTD/Totalizador (ver Secção 2) para o ano seguinte, com uma granularidade de 15 minutos, de modo a que o objetivo final do trabalho, visando a validação de dados para colmatar falhas de registo, pudesse ser cumprido.

A construção dos perfis para o ano seguinte (Novembro de 2016 a Outubro de 2017) foi realizada em R. As leituras efetuaram-se por cada PTD_ID (CodigoPTD e Totalizador).

Os elementos de entrada utilizados nesta etapa foram:

- os perfis ou séries temporais correspondentes aos centróides obtidos no *Clustering* pelo *K-Means,* calculados a partir dos valores de potência correspondentes aos consumos nos PTD_ID's, valores esses que tinham sido normalizados para evitar a influência da heterogeneidade dos valores entre os vários PTD's no algoritmo de *Clustering* (*K-Means*);
- os *clusters* a que pertence cada PTD_ID, que foram obtidos, no caso dos PTD's com dados completos, pelo *Clustering*, e no dos PTD's com dados incompletos, pela Classificação com o algoritmo *Random Forest*;
- a soma de todas as potências lidas num mês (potência total) e o número de variáveis temporais distintas (períodos de 15 minutos) a que correspondem essas leituras, por cada PTD_ID.

A partir destes *inputs,* foram efetuados os seguintes procedimentos:

1. construiu-se uma tabela com uma coluna com todas as variáveis temporais (data e hora) do período a prever e as restantes colunas, uma

por *cluster*, contendo as coordenadas dos centróides desses *clusters*. Para isso houve necessidade de *desagrupar* as coordenadas dos clusters que se encontravam agrupadas por mês, dia da semana e hora na granularidade inicial (de 15 em 15 minutos);

2. construiu-se uma tabela com uma linha por cada PTD_ID contendo as coordenadas do centróide do *cluster* que lhe corresponde, *desagrupadas* como se descreveu no ponto anterior. Cada coluna contém as coordenadas correspondentes a uma variável temporal (data e hora) exceto uma coluna que contém o PTD_ID, outra com a potência total (soma das potências) desse PTD_ID e outra com o número de variáveis temporais (data e hora) registadas a que a potência total corresponde;

3. calculou-se o perfil das potências previstas escalando de forma proporcional (multiplicando por uma constante de proporcionalidade ou fator de escala) os valores das coordenadas de cada centróide por PTD_ID. Adotou-se, como constante de proporcionalidade, o quociente das potências médias por período de 15 minuto (soma das potências a dividir pelo número de leituras (variáveis temporais distintas lidas) pela média das coordenadas de potência do centróide. Assim, para cada PTD_ID $\boldsymbol{P}$ pertencente ao *cluster* **c** e para uma dada variável temporal (data/hora) $\boldsymbol{i}$ a potência prevista é dada por:

$$\text{potência prevista}(P) = \frac{\text{potência média quarto-horária}}{\text{média}(c)} \times c_i$$

onde $\boldsymbol{c_i}$ é a coordenada $\boldsymbol{i}$ do cluster $\boldsymbol{c}$.

Obtém-se após este procedimento uma tabela no R com o PTD_ID, o número do *cluster* a que pertence cada PTD_ID, potência média por leitura, de cada mês e, finalmente, os valores de potência previstos pelo modelo que seriam registados em todos os intervalos de 15 minutos, como exemplificado na Tabela

2. Devido às dificuldades de processamento no R pela quantidade de dados a processar, este procedimento foi realizado para cada mês individualmente. Assim, por exemplo para o mês de Novembro, as restantes colunas da tabela referida correspondem às previsões dos valores de potência em cada 15 minutos desde as 00h00 de 1 de Novembro de 2016 até às 23h45 do dia 30 de Novembro de 2016, período a definir no início do procedimento a cada utilização.

Tabela 2

Esquema do *data.table* obtido para efetuar as previsões para o período de Novembro de 2016 a Outubro de 2017.

PTD_ID	Cluster	Consumo Médio Mensal	t1	t2	…	tn
…	…	…	Previsões para o período de Novembro de 2016 a Outubro de 2017			
…	…	…				
…	…	…				
…	…	…				

6.7. Análise dos Resultados

A última etapa deste trabalho consistiu no cálculo do erro percentual absoluto entre valores reais e previsões, para avaliar a precisão das previsões obtidas a partir do *Clustering* e da Classificação.

Depois de realizada a previsão para o período considerado, foi calculado o erro percentual absoluto entre as previsões e os valores reais, através da seguinte expressão:

$$\frac{|previsão - valor\ real|}{valor\ real} \text{ x } 100\%$$

Dado que estavam apenas disponíveis dados reais de Novembro de 2016, os cálculos do erro foram realizados para avaliar apenas a precisão das previsões para um só mês. A comparação foi realizada em R com o intuito de compreender se as nossas previsões eram representativas da realidade. Para

este efeito foram importados dados reais de Novembro de 2016 de 15 em 15 minutos, assim como as previsões que resultaram da Secção 6.7. para o período de Novembro de 2016. Foi inicialmente calculado o erro para cada PTD_ID e, finalmente, uma média desses erros. Logicamente, foram apenas utilizados para o cálculo do erro PTD's presentes em Novembro de 2015, que contém os dados que estão na origem das previsões, e em Novembro de 2016, que contém os dados reais. Foram também realizados histogramas para avaliar a distribuição do erro. Devido à existência de possíveis *outliers,* os erros que se encontravam fora do intervalo definido pela média mais dois desvios padrão foram retirados e o erro médio foi novamente calculado.

7. Resultados e Discussão

7.1. Estimativas de potência, inferência de *missing values* e Normalização de dados

A primeira etapa do trabalho consistiu em fazer uma análise dos métodos de estimativa a utilizar para o cálculo de valores de potência em falta e dos erros associados aos métodos selecionados, tendo em conta fatores como o número de pontos de estimativa, a duração das falhas e o tamanho da amostra utilizada para o estudo. Como referido anteriormente, apenas dois métodos foram considerados: média e regressão linear. Estes métodos foram escolhidos pela sua simplicidade e porque são utilizados no tratamento de dados no sector energético(Chen, Li, Lau, Cao, & Wang, 2010).

A Figura 3 apresenta as variações do erro percentual absoluto global consoante o método utilizado, o tamanho da amostra utilizada para testar o método, o número de pontos de estimativa e a duração das falhas. Nos gráficos, cada linha corresponde a uma amostra de simulação da falha (N=60, 600, 6000 ou 60000), onde o erro varia com o aumento do número de pontos de estimativa e da duração das falhas. Os valores 1, 2 e 3 no eixo das abcissas correspondem ao número de pontos de estimativa que foram utilizados para calcular a estimativa. Analisando os resultando obtidos, foi possível compreender que existem maiores variações do erro percentual absoluto global utilizando a média e que, de uma forma geral, à medida que aumenta a duração da falha, o erro aumenta, o que faz todo o sentido já que estimar um conjunto de valores maior será menos preciso.

A Tabela 3 apresenta uma análise mais detalhada, da qual excluímos N=60 por apresentar resultados incertos e ser uma amostra de tamanho muito reduzido, que estatisticamente apresenta resultados menos precisos. De acordo com esta tabela, para N=600, 6000 e 60000, o erro aumenta à medida que a duração da falha aumenta, à exceção de dois casos: N=600 em que o erro

diminui dos períodos de 45 para os de 60 minutos e N=60000 em que o erro diminui dos de 30 para os de 45 minutos. Não foi possível encontrar explicações para este facto à exceção da aleatoriedade das amostras extraídas para simular as falhas. Cada linha corresponde a uma amostra de simulação da falha onde o erro com o aumento do número de pontos de estimativa e da duração das falhas.

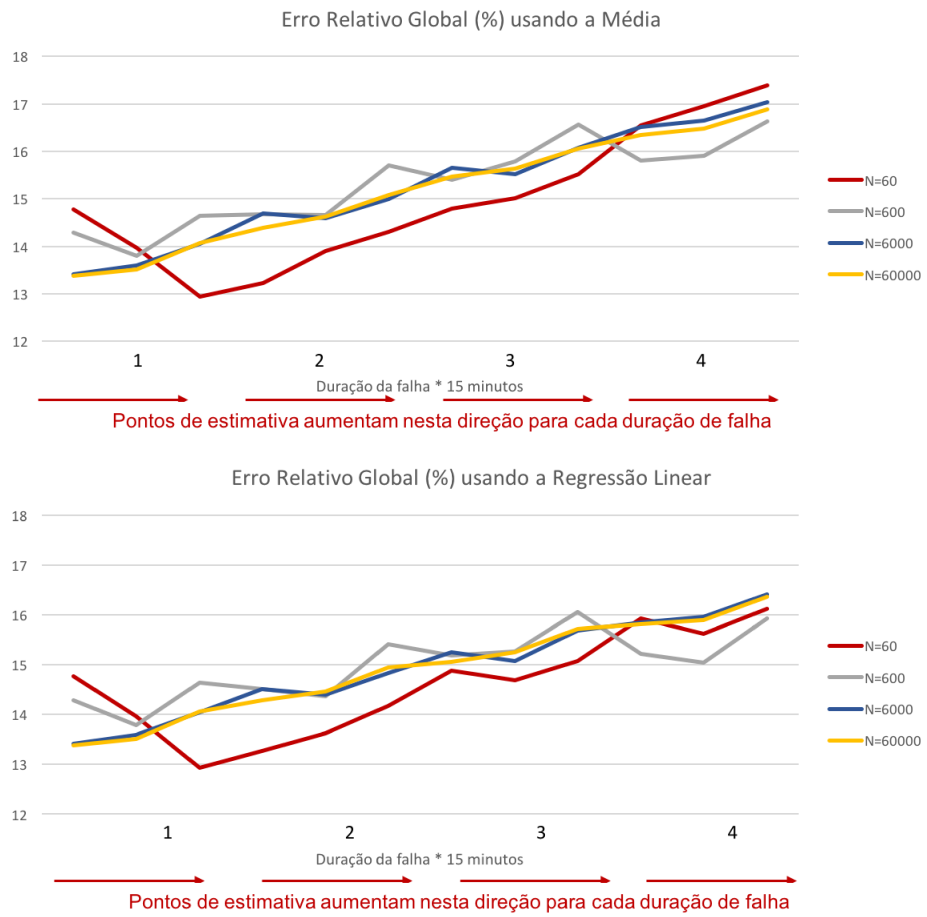


Figura 3

Erros Absolutos Percentuais Globais (%) obtidos através da Média e da Regressão Linear para as dimensões da amostra aleatória N=60, 600, 6000, 60000 para duração de falha de 15, 30, 45 e 60 minutos, utilizando 1,2 e 3 pontos de estimativa.

Tabela 3

Erros Absolutos Percentuais Globais (%) obtidos através da Média e da Regressão Linear para as dimensões da amostra aleatória N=60, 600, 6000 e 60000, para duração de falha de 15, 30, 45 e 60 minutos, utilizando 1,2 e 3 pontos de estimativa.

Duração da Falha *15 minutos	Nº Pontos de Estimativa	MÉDIA				REGRESSÃO LINEAR			
		N=60	N=600	N=6000	N=60000	N=60	N=600	N=6000	N=60000
1	1	14,770	14,288	13,419	13,377	14,770	14,288	13,419	13,377
1	2	13,966	13,791	13,592	13,510	13,966	13,791	13,592	13,510
1	3	12,936	14,634	14,050	14,061	12,936	14,634	14,050	14,061
2	1	13,228	14,675	14,700	14,395	13,274	14,505	14,510	14,289
2	2	13,896	14,656	14,585	14,621	13,628	14,369	14,404	14,464
2	3	14,309	15,707	15,004	15,072	14,174	15,409	14,837	14,944
3	1	14,797	15,403	15,653	15,461	14,875	15,188	15,244	15,065
3	2	15,013	15,796	15,515	15,632	14,681	15,260	15,075	15,254
3	3	15,516	16,561	16,083	16,052	15,067	16,057	15,691	15,724
4	1	16,549	15,811	16,510	16,341	15,928	15,224	15,852	15,816
4	2	16,956	15,908	16,656	16,486	15,612	15,037	15,963	15,896
4	3	17,385	16,623	17,035	16,876	16,113	15,922	16,411	16,362

A Figura 4 apresenta uma análise mais detalhada dos erros nas várias situações.

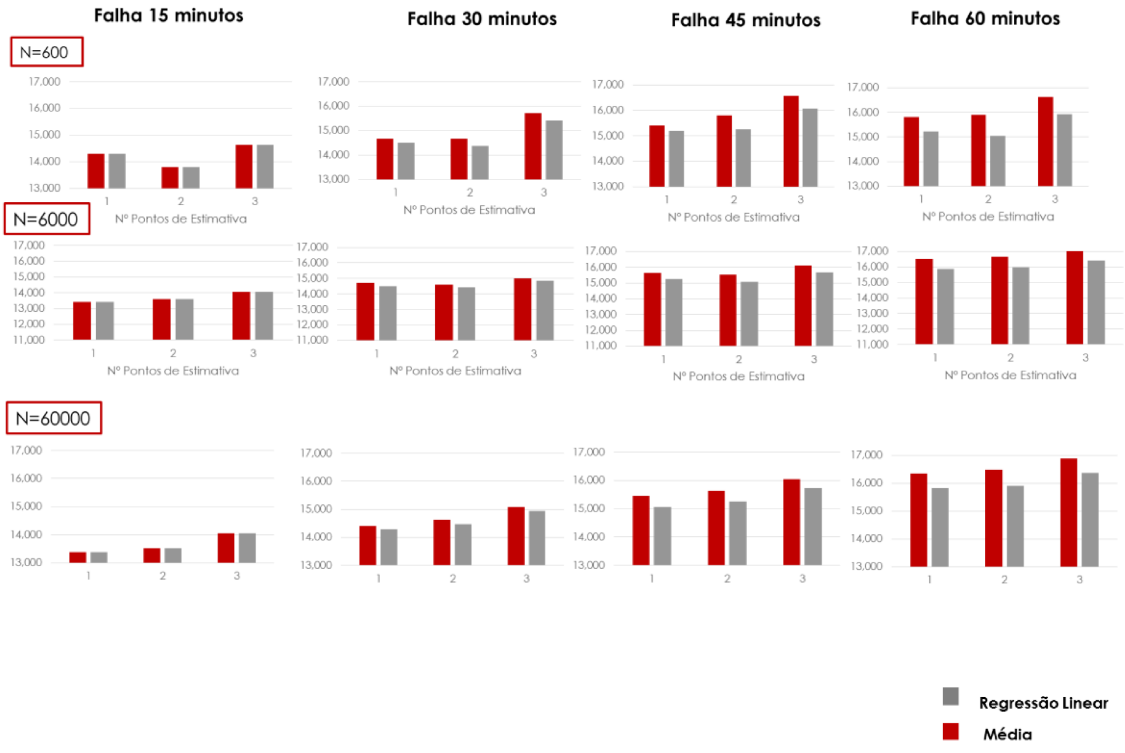


Figura 4

Erros Absolutos Percentuais Globais (%) obtidos através da Média e da Regressão Linear para as dimensões da amostra N=600, 6000 e 60000, para duração de falha de 15, 30, 45 e 60 minutos, utilizando 1,2 e 3 pontos de estimativa.

Com base nesta figura, conclui-se que o método que apresenta menor erro é a regressão linear (a cinzento) em todos os casos. Nas falhas de 15 minutos é indiferente utilizar um ou outro método, já que o resultado é o mesmo. De facto era de esperar que a regressão linear constituísse um método melhor pois considera a tendência da série temporal, enquanto a média resulta no mesmo valor inferido para todos os instantes com registos em falta, independentemente da potencial existência de tendência na série.

Relativamente ao número de pontos de estimativa a utilizar, e considerando a Figura 4, concluiu-se que para N=600 e N=6000 os erros são menores quando são considerados 2 pontos de estimativa enquanto para N=60000, os erros são menores quando é considerado apenas 1 ponto antes e depois da falha. Perante a indecisão de qual o número de pontos de estimativa a utilizar, a decisão final foi complementada com uma análise geral do erro à medida que o tamanho da amostra aleatória utilizada aumentava. Na Tabela 4, é possível ver que o erro vai estabilizando, já que a diferença entre N=600 e N=6000 pode atingir os 0,9%, mas a diferença entre N=6000 e 60000 é menor do que 0,2%. Assim sendo tornou-se mais sensato utilizar dois pontos de estimativa dado que contribuem com mais informação do que um só ponto.

Tabela 4

Erros Absolutos Percentuais Globais (%) obtidos através Regressão Linear para as dimensões da amostra 600, 6000 e 60000 e para as dimensões para duração de falha de 15, 30, 45 e 60 minutos, utilizando 2 pontos de estimativa. Comparação dos Erros Absolutos Percentuais Globais em percentagem à medida que aumenta a dimensão da amostra.

Regressão Linear (2 ponto de estimativa)			Diferenças
Duração da falha	N=600	N=6000	
15 minutos	13,791	13,592	0,200
30 minutos	14,369	14,404	0,034
45 minutos	15,260	15,075	0,185
60 minutos	15,037	15,963	0,927

Regressão Linear (2 ponto de estimativa)			Diferenças
Duração da falha	N=6000	N=60000	
15 minutos	13,592	13,510	0,082
30 minutos	14,404	14,464	0,060
45 minutos	15,075	15,254	0,179
60 minutos	15,963	15,896	0,067

Finalmente, com o intuito de comparar o erro global com o erro por período, avaliaram-se os erros nas duas situações (Tabela 5), tendo concluído que os períodos Manhã I e Manhã II apresentavam as maiores diferenças relativamente ao erro global, mas que, apesar disso, estas não pareciam ser significativas para que o cálculo de estimativas tivesse que ser realizado por período. Ainda relativamente aos períodos, convém referir que estes não foram criados de forma totalmente aleatória, mas foram estabelecidos com base no conhecimento do negócio, considerando que alguns períodos do dia poderiam apresentar maiores similaridades devido ao comportamento diário do setor residencial, cuja energia é fornecida pelos PTD's.

Tabela 5

Comparação entre os Erros Absolutos Percentuais por Período (%) e o Erro Absolutos Percentuais Global (%) obtidos através do método de Regressão Linear, para durações de falha de 15, 30, 45 e 60 minutos, utilizando 2 pontos de estimativa.

	Regressão Linear (2 pontos de estimativa)						
Duração da falha	Erro Global	Erro Manhã I	Erro Manhã II	Erro Meio do Dia	Erro Tarde	Erro Resto do Dia	Erro Noite
15 minutos	13,3770	15,844	14,385	14,156	13,741	11,435	10,829
30 minutos	14,2893	16,499	15,912	15,317	14,359	12,333	11,459
45 minutos	15,0646	17,585	16,660	16,259	15,185	13,367	11,483
60 minutos	15,8161	18,449	17,545	17,073	15,949	14,379	11,662

Assim sendo, o método utilizado para estimar valores de potência correspondentes a falhas não superiores a uma hora foi a regressão linear, utilizando dois pontos de estimativa (2 registos existentes de cada lado da falha).

A etapa de preenchimento de falhas foi realizada nesta fase como tentativa de reduzir os *missing values* presentes na base de dados e para possibilitar a utilização de um número maior de PTD's para a fase de *Clustering*. O preenchimento das falhas não foi realizado para períodos superiores a uma hora, já que o erro associado à estimativa ultrapassava os 15%. Optou-se então por considerar que a partir desse valor, as estimativas já não eram tão precisas e não deveriam ser realizadas.

É de sublinhar também que as potências foram normalizadas recorrendo a um método de normalização que utiliza os valores normalizados dos extremos da escala para que, aquando da AED ou do *Clustering*, as diferenças na magnitude de valores entre PTD's não interferissem. Esta etapa é de extrema importância no modo como se desenvolveu este trabalho, já que, a partir do momento em que os dados foram normalizados em função dos valores de potência para cada PTD, passou-se a considerar apenas a forma dos perfis e não a escala. Ou seja, aquando da realização do *Clustering*, dois PTD's com consumos diferentes (escala) mas com variações de potência ao longo do ano semelhantes (forma) iriam ser considerados como pertencendo a um mesmo *cluster*. Ou seja, a segmentação iria ser realizada tendo em conta apenas a forma dos perfis.

7.2. Análise Exploratória dos Dados

I) Nesta etapa da análise exploratória dos dados, não foi utilizada a totalidade dos registos fornecidos. Convém referir que esta primeira etapa da AED foi realizada utilizando as potências não normalizadas. Optou-se por realizar uma análise, Dia versus Noite, Dias Úteis vs. Fins-de-Semana ao longo da semana e ao longo do mês, que evidenciasse que é possível segmentar os PTD's manualmente. Esta análise incluiu apenas 1000 PTD's do mês de Julho (escolhidos aleatoriamente) para facilitar a visualização dos dados.

No gráfico da Figura 5, observam-se PTD's com características muito distintas ao longo do dia, nomeadamente entre as 8 e as 21 horas, ou seja, é provável que um algoritmo de *Clustering* seja capaz de detetar essas diferenças.

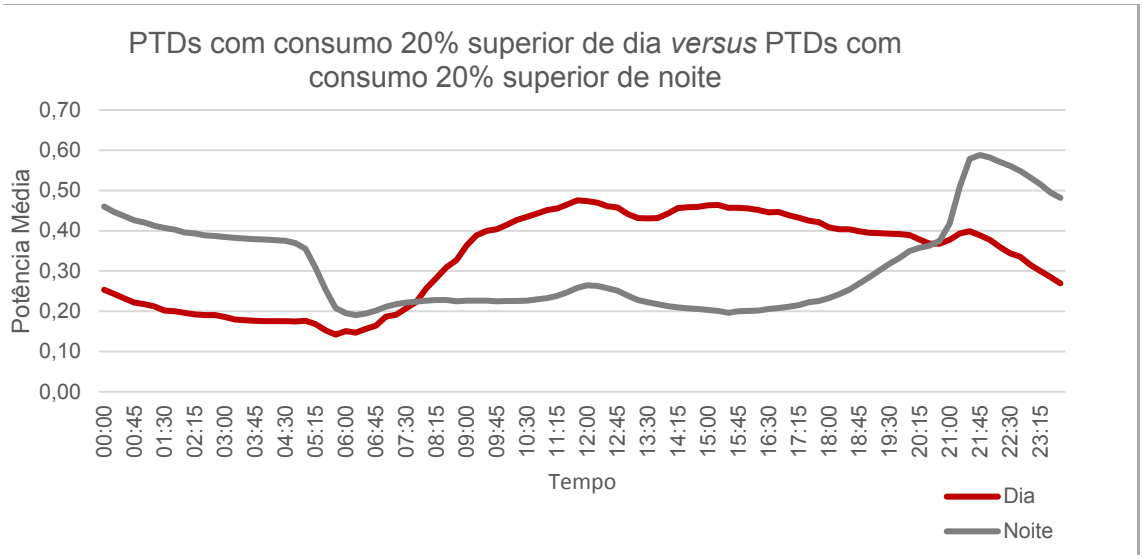


Figura 5

Análise Dia versus Noite de uma amostra de 1000PTDs do mês de Julho.

Do mesmo modo, no gráfico da Figura 6, é possível visualizar perfis muito diferentes entre PTD's com um consumo 20% maior em dias úteis e PTD's com um consumo 20% maior no fim de semana. No entanto, visto que as diferenças entre perfis ao longo do mês, presentes na Figura 5 não são percetíveis de forma clara, esta mesma análise foi realizada considerando que há uma semana-tipo característica do mês (Figura 6).

A Figura 7 mostra também que para PTD's com um consumo maior em dias úteis do que no fim de semana, existe uma variação de aproximadamente 0,4 de amplitude (valor normalizado de potência) ao longo dos dias úteis e que, durante o fim de semana, essa amplitude diminui aproximadamente para metade como seria de esperar se se considerar que uma grande parte destes PTD's fornecem energia ao setor residencial.

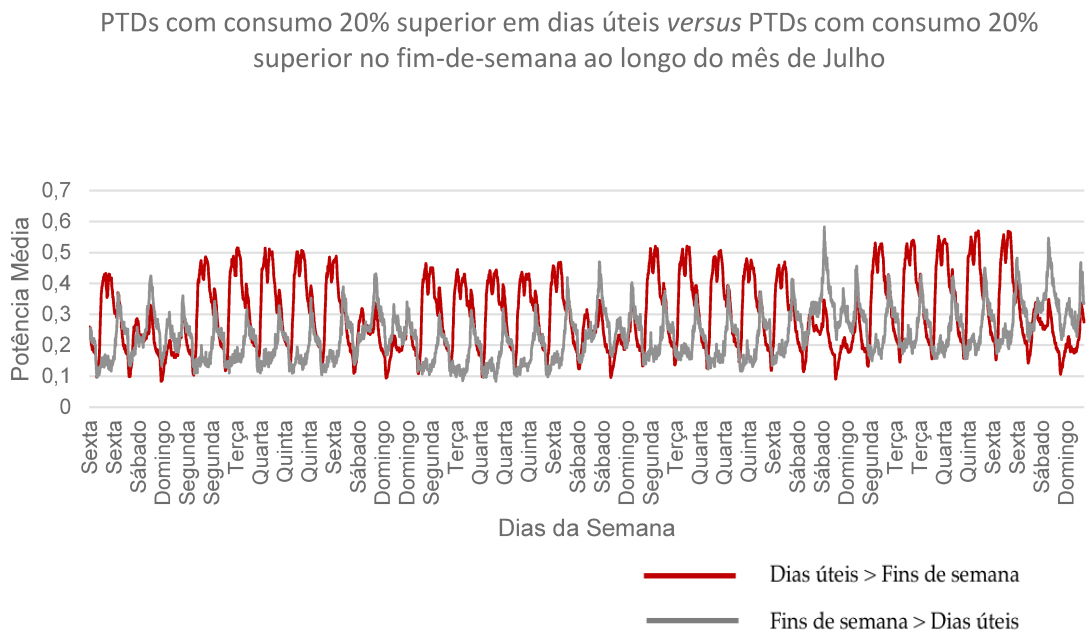


Figura 6

Análise Dias Úteis versus Fim de semana de uma amostra de 1000PTDs do mês de Julho, ao longo do mês.

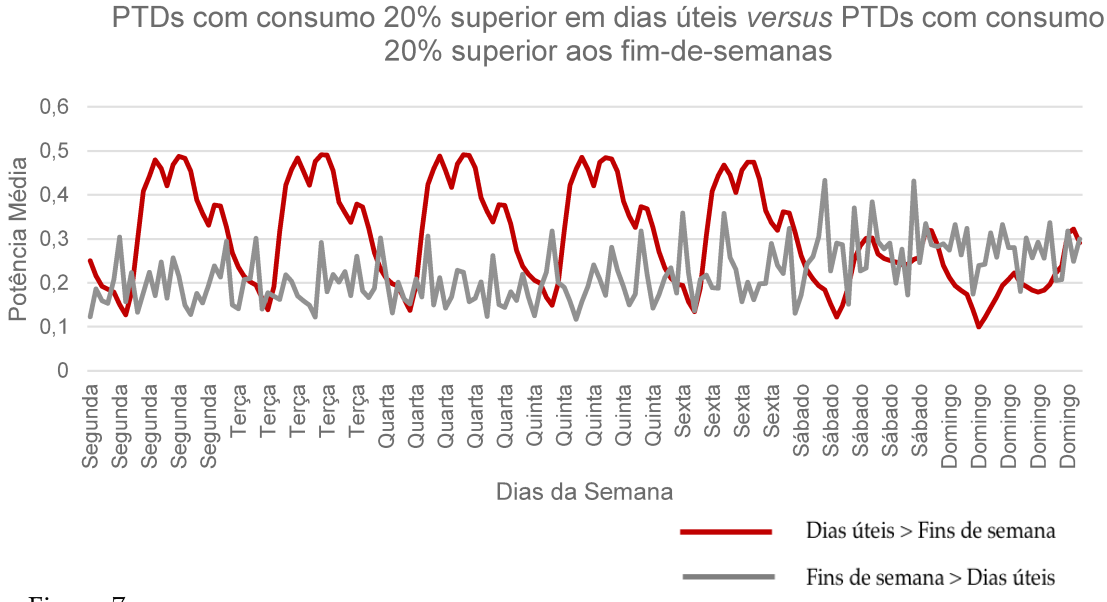


Figura 7

Análise Dias Úteis versus Fim de semana de uma amostra de 1000PTDs do mês de Julho, considerando dias da semana tipo durante o mês.

II) A segunda etapa da AED consistiu em reduzir as variáveis no horizonte temporal, para facilitar o processamento dos dados no *Clustering*. Convém

referir que nesta etapa, as potências já estavam normalizadas de acordo com o ponto 7.1.

Para melhor perceção da dimensão que estas variáveis poderão atingir, convém analisar a estrutura dos dados extraídos do SQL Server e a estrutura modificada dos dados com que se efetuou o *Clustering* (Figura 2 da Secção Metodologia).

Tendo em vista a redução do número de variáveis temporais a utilizar no *Clustering*, foram realizados testes para averiguar se era possível agrupar os dados, utilizando medidas agregadoras como a média. Os testes de agrupamento dos valores de potência foram feitos ao nível das horas, do dia e do dia da semana tipo em cada mês. Esta etapa foi extremamente relevante para contornar situações em que o R não consegue processar um volume tão grande de dados, devido ao espaço disponibilizado pela memória RAM.

O teste que se efetuou baseou-se na similaridade entre vetores, por intermédio do cálculo do cosseno entre os mesmos. Os resultados do teste do cosseno entre os vetores *v* e *vt* sugeriram que é possível agrupar por hora e por dia tipo da semana. Obtiveram-se valores de 95% e 89% de registos com valor de cosseno entre os vetores *v* e *vt* superior a 0,9 (Tabela 9).

Tabela 6
Resultados dos testes do cosseno.

Tipo de Agrupamento	**% Registos com cos α > 0,9**
Hora	95%
Dia	44%
Dia da Semana Tipo	89%

Perante estes resultados, conseguiu-se reduzir significativamente o número de variáveis temporais consideradas, de 96 registos diários x 365 dias = 35040 num ano, para 24 registos x 7 dias da semana tipo x 12 meses = 2016. Isto facilitou a importação dos dados e posterior processamento no R.

Convém referir que, antes de realizar o teste do cosseno, foi realizada uma tentativa de reduzir as variáveis por intermédio de um *t-test,* no sentido de verificar se haveriam diferenças significativas entre os registos agregados com base em diferentes horizontes temporais.

Os resultados dos testes revelaram que a hipótese nula era particamente sempre rejeitada, ou seja, rejeitava a hipótese de que os registos eram semelhantes. Esta é uma limitação dos testes de hipóteses quando aplicados a amostras de grandes dimensões, uma vez que o valor da estatística de teste recorre à raiz quadrada do número de observações no denominador, fazendo com que o valor da estatística do teste seja praticamente nula, rejeitando a hipótese nula (Anderson, Burnham, & Thompson, 2000).

7.3. *Clustering*

Para melhor compreender a descrição dos resultados que se seguem, é necessário compreender, em primeiro lugar, quais as considerações realizadas para a escolha do *software* e do algoritmo para a realização do *Clustering* e a forma como os desafios que foram surgindo foram encarados, nomeadamente o grande volume de dados. Relativamente ao *software* a utilizar, foi escolhido o R pela variedade de técnicas de *Clustering* que oferece, pela rapidez com que executa os processos comparativamente com outros *softwares* e pelo conhecimento do *software* existente. No entanto, a utilização de *software* ou ferramentas, como o R, com grandes volumes de dados, pode-se revelar um desafio, já que o R utiliza a memória RAM para armazenar dados aquando do processamento. Ainda, na escolha dos métodos de C*lustering* a utilizar, existem várias questões a considerar, além do software: o tamanho (número de observações), a dimensionalidade (número de variáveis) dos dados e a existência de *missing values*. Um outro aspeto a considerar no caso dos diagramas de carga, será o facto de estes se apresentarem segundo uma série temporal. A escolha da técnica a utilizar para a segmentação prendeu-se então com estas considerações.

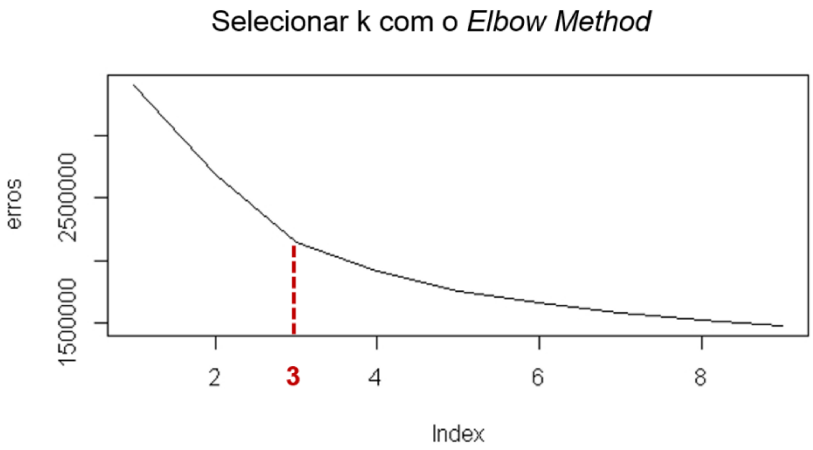
Relativamente à existência de *missing values* não estimados na primeira etapa (relativos a falhas superiores a 1h), estes foram desconsiderados para a análise de *Clustering* (10442/68500 * 100 = cerca de 15%).

Para lidar com um dos maiores problemas acima referidos, a total alocação da memória RAM utilizada pelo R, várias tentativas foram realizadas no sentido de criar sensibilidade relativamente à quantidade de dados que podiam ser importados para o R. Após algumas tentativas, percebe-se que se conseguia apenas importar apenas 2 meses de dados (um de cada vez), no caso de os registos estarem agrupados por hora, mas que se conseguia importar um ano inteiro de dados se estivessem agrupados por hora e por dia da semana tipo. Como já referido anteriormente, considerou-se que os valores de potência normalizados podiam ser agrupados deste modo. Assim sendo, e tendo em conta que se pretendia realizar um *Clustering* com os dados de um ano, procedeu-se à exportação dos dados agrupados do SQL Server para o R. Os dados foram modificados previamente ao *Clustering*, tal como descrito na Figura 2 da Secção Metodologia, para facilitarem a aplicação do algoritmo.

A segmentação foi realizada utilizando o método particional K-Means, que se baseia na distância euclidiana, em detrimento do método *dtwclust*, que se baseia no DTW. A razão pela qual isto aconteceu está relacionada com o facto do processamento utilizando o *dtwclust* com o mesmo volume de dados ultrapassar as 120horas de processamento. No entanto, convém referir que o *dtwclust* poderia ser mais indicado nesta fase, no lugar da distância euclidiana, para que se tornasse possível agrupar séries que estão desalinhadas apenas uma hora no mesmo *cluster*. No entanto, dado que a granularidade das séries temporais já não é tão fina como nos dados extraídos dos diagramas de carga dos PTD's, deixa de ser tão importante considerar os desvios, até porque desalinhamentos superiores a uma hora poderão já revelar variações demasiado grandes.

Para a utilização do algoritmo K-Means, é necessário definir o número de *clusters*. Com esse intuito recorreu-se ao índice de Davies-Bouldin (DB índex), sendo que quanto menor o valor, mais adequado o k, e ao *Elbow Method*,

baseado na construção de uma curva, que avalia a variância explicada em função do número de *clusters* escolhidos (Figura 8).



Nº Clusters	DB Index
1	NaN
2	1,851
3	**1,504**
4	1,708
5	1,690
6	1,676
7	1,703

Figura 8

Resultados do número de *clusters* ótimo a utilizar para a segmentação com o Método de Elbow e o DB índex.

Ambos determinaram que k=3 era o número ótimo de *clusters* a considerar (Figura 8), resultado que se demonstrou algo surpreendente já que existe uma grande variedade de PTD's. No entanto, dado que os dois métodos estavam em concordância, optamos por avançar com este número de *clusters*. Os resultados são visíveis na Figura abaixo.

Antes da análise dos resultados do *Clustering*, convém referir que foi realizado um estudo para avaliar se os resultados anuais obtidos (de Novembro de 2015 a Outubro de 2016) são representativos de cada mês ou não. Foram escolhidos dois meses (Abril e Setembro), de forma aleatória, que estivessem afastados entre si no tempo, com o intuito de observar se estes apresentavam comportamentos diferentes relativamente ao ano, e foram realizadas análises de *Clustering* considerando cada um dos meses. Os resultados contidos na Tabela 7 mostram os PTD's com idêntica segmentação entre Abril e o ano, e entre Setembro e o ano. Se calcularmos a percentagem de PTD's comuns em cada mês relativamente ao ano, pode-se verificar que:

- De 82% a 92% dos PTD's presentes em Setembro de 2016 estão juntos nos mesmos *clusters* no período anual;
- De 87% a 93% dos PTD's presentes em Abril de 2016 estão juntos nos mesmos *clusters* no período anual.

Tabela 7

Análise Dias Úteis versus Fim-de-Semana de uma amostra de 1000PTDs do mês de Julho, considerando dias da semana tipo durante o mês.

Número de PTD's em comum entre os clusters de Setembro e Abril de 2016 e os clusters anuais

	Datas	Nov'15 - Out'16		
Datas	Cluster	1	2	3
	Cluster	Número PTD's	6870	17048	34140
Set'16	1	19561		**15681**	
	2	35040			**29755**
	3	9793	**5605**		
Abr'16	1	38302			**6025**
	2	18389		**15658**	
	3	9124	**31868**		

Estes resultados reforçam a ideia de que o perfil do ano é representativo do dos meses.

Os resultados do *Clustering* apresentados seguidamente foram obtidos em R mas analisados mais detalhadamente no Microsoft Office Excel. Os perfis apresentados nos gráficos dizem respeito aos 3 perfis dos centróides obtidos no período considerado para a segmentação (Figura 9).

Analisando os resultados do *Clustering*, pode-se considerar que foram obtidos 3 *clusters* com um comportamento diferente. Numa análise global dos resultados, verifica-se que o consumo dos PTD's do *cluster1* apresenta menor amplitude, quando comparado com os outros dois. Apesar de não ser muito percetível nos *clusters* 1 e 3, pode-se verificar no *cluster* 2 o mesmo padrão de variação ao longo dos 12 meses. Convém referir que existe um aumento de consumo nos meses de Verão (Julho, Junho e Agosto) nos 3 casos, de onde se pode concluir que no Verão o consumo de energia é maior.

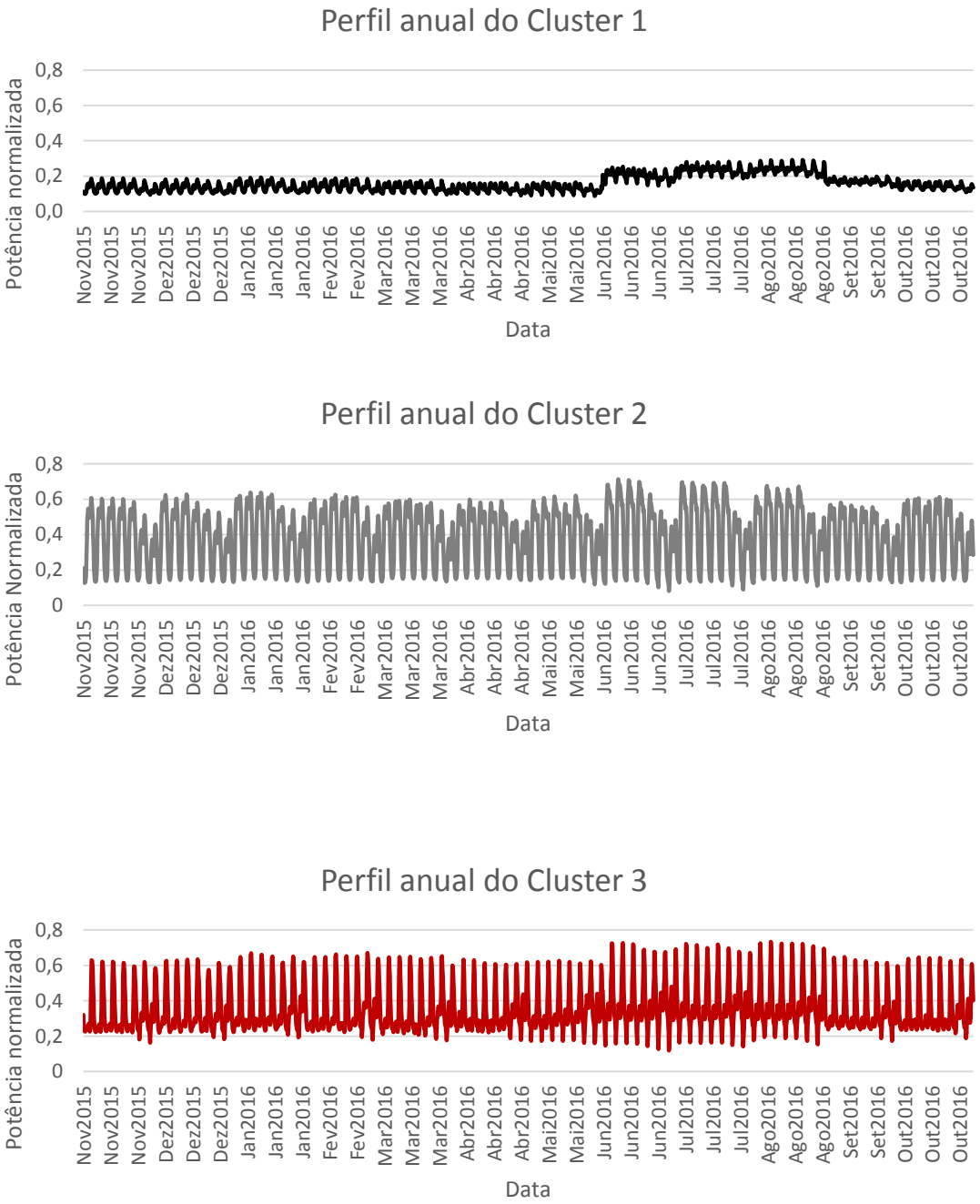


Figura 9

Resultados do *Clustering:* perfis dos centróides para o período de Novembro de 2015 a Outubro de 2016.

Para fazer uma análise comparativa e mais detalhada do comportamento dos PTD's, foi escolhido o mês de Outubro de 2016, apenas por ser o mês mais

recente de entre o período selecionado. A Figura 10 abaixo apresenta os perfis do mês de Outubro dos 3 centróides obtidos a partir do *Clustering*.

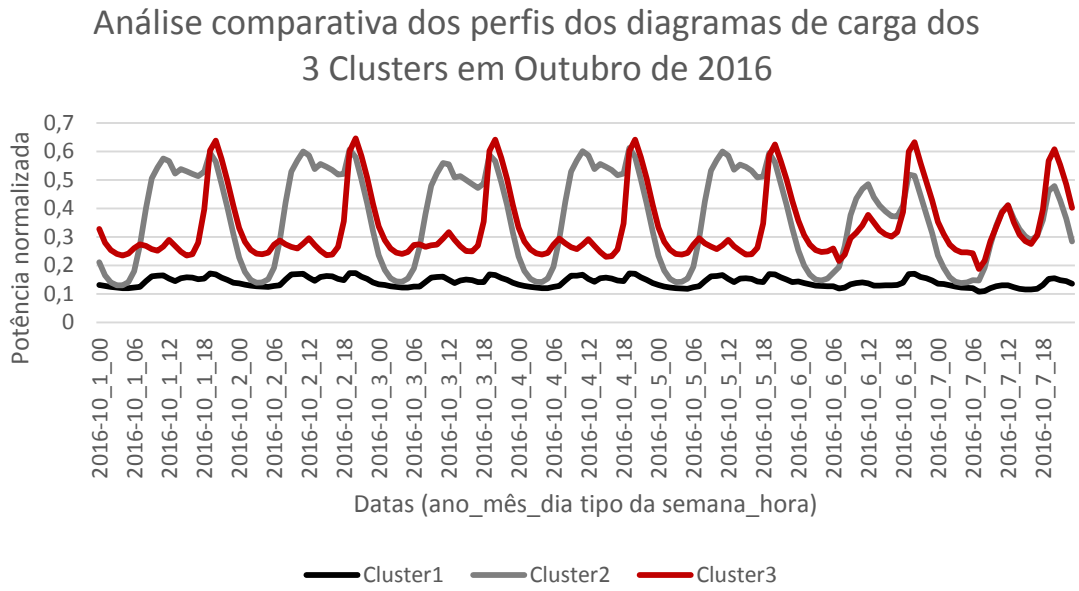


Figura 10

Resultados do *Clustering:* perfis dos centróides de Outubro de 2016.

Analisando gráfico, conclui-se, mais uma vez, que a amplitude do *cluster* 1 é tão reduzida comparativamente aos outros que se torna difícil avaliar o seu comportamento individual. No entanto, considerando o *cluster* 2, é possível observar um aumento da potência significativo a partir das 6 horas dos dias, consumo este que se mantém sempre elevado até à meia noite, momento em que sofre uma diminuição abrupta. Olhando para o *cluster* 3, é possível verificar que o consumo às 6 horas não é tão baixo como o do *cluster* 1, e não é tão baixo como o consumo mínimo do *cluster* 2. Neste terceiro *cluster*, o consumo mantém-se ligeiramente constante até às 18horas, a partir das quais sofre um aumento considerável até a uma potência normalizada de 0,6 aproximadamente.

Na Figura 11 também conseguimos observar ligeiras diferenças entre dias úteis e o fim-de-semana, mais concretamente dias tipo 5, 6 e 7 da semana. Para uma análise destas diferenças, foram traçados gráficos entre quintas-feiras e domingos dos 3 *clusters.*

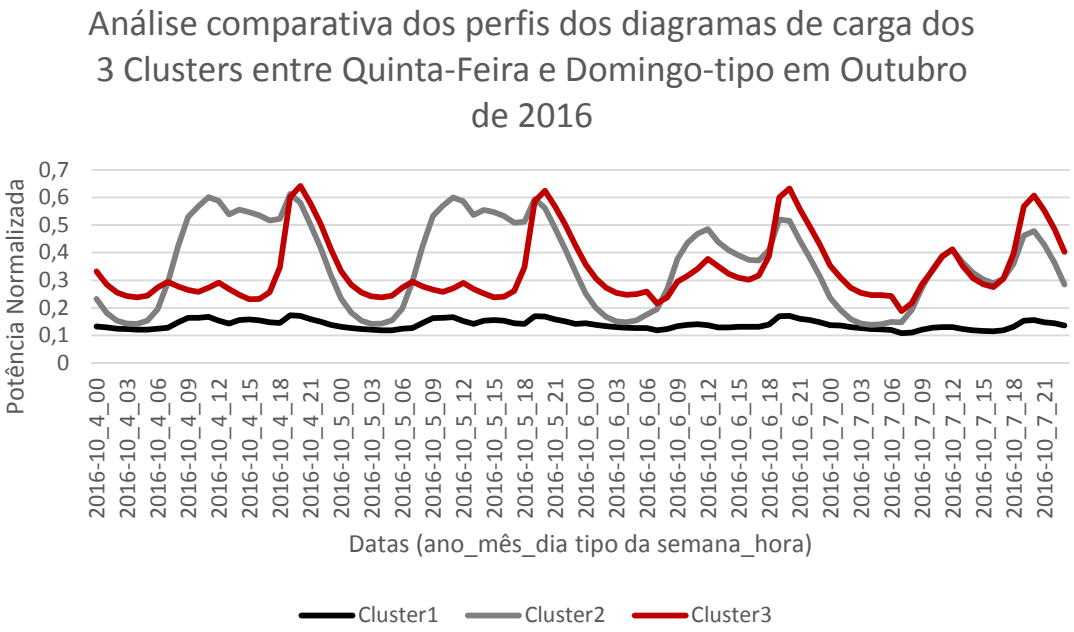


Figura 11

Resultados do *Clustering:* perfis dos centróides entre quinta-feira tipo e domingo-tipo de Outubro de 2016.

Neste gráfico é possível verificar, relativamente ao *cluster* 2, que há uma diminuição do consumo máximo de Sexta para Sábado e de Sábado para Domingo. Relativamente ao *cluster* 3, este apresenta um consumo máximo que se mantém no fim-de-semana e um aumento de consumo entre as 6 e as 15 horas de Sábado e Domingo. Dado que o *cluster* 1 apresenta dados aparentemente constantes quando comparado com os outros *clusters*, procedeu-se à análise de modo individual para compreender o seu comportamento em dias úteis e fins-de-semana (Figura 12).

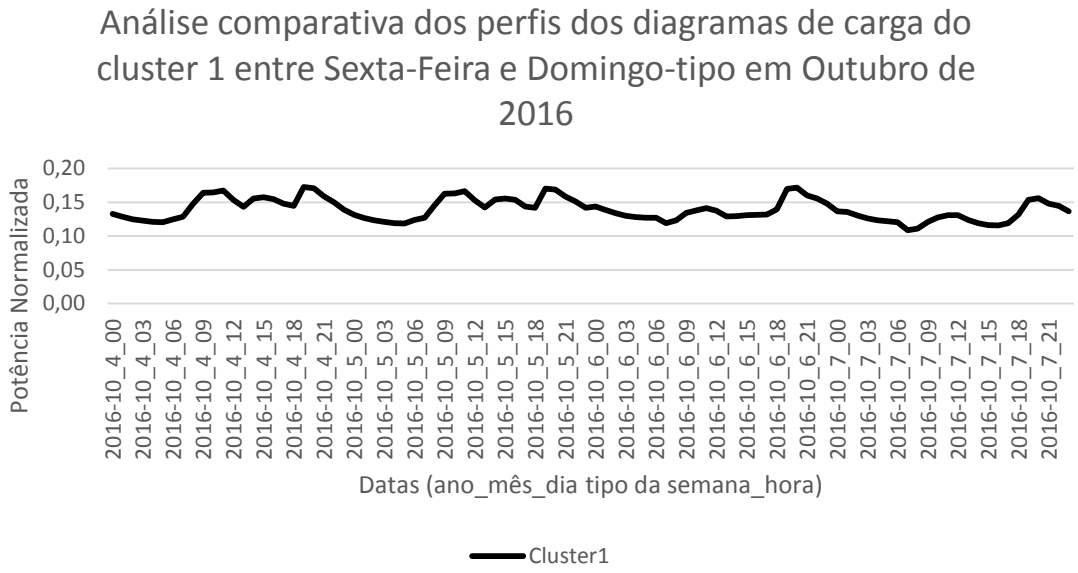


Figura 12

Resultados do *Clustering:* perfis do centróide correspondente ao Cluster1 entre sexta-feira tipo e domingo-tipo de Outubro de 2016.

Olhando para o gráfico, pode-se verificar que as variações continuam a não ser muito significativas, não havendo um padrão óbvio de evolução e concluiu-se que, de um modo geral, o consumo parece ser ligeiramente menor ao fim-de-semana e que da meia-noite às 6 horas o consumo é menor do que nas restantes horas.

7.4. Classificação

A Classificação surgiu não só como uma forma de minimizar possíveis problemas na realização do *Clustering*, mas também como forma de criar um modo de classificar PTD's que poderão vir a ser instalados no futuro e PTD's com *missing values*. Após o *Clustering*, o objetivo passou por alocar esses PTD's aos *clusters* formados, assumindo que nos locais onde existem falhas, os PTD's assumiriam os valores correspondentes às estimativas efetuadas.

Assim, a pesquisa realizada refletiu-se em:

- Classificar séries temporais descontínuas, já que, no caso dos PTD's que foram excluídos pela presença de *missing values*, o algoritmo de classificação terá de ser capaz de classificar os PTD's apenas com base nas variáveis ou registos temporais existentes;
- Classificar séries temporais parciais, no caso dos PTD's que possuem dados inferiores ao período considerado para a realização do *Clustering*.

O número de PTD's (CódigoPTD+Totalizador) que faltavam alocar aos *clusters* existentes era de 10442. Dada a heterogeneidade do posicionamento das falhas presentes nestes PTD's e da duração das mesmas, foi necessário desenvolver um modelo de classificação para cada um dos PTD's, sendo que cada modelo considerava apenas as variáveis existentes do PTD a classificar. O facto de esta metodologia ser tão específica e consumir muito tempo de processamento, remeteu para que se considerasse outra técnica para a classificação.

Aquando da Classificação foram consideradas então duas metodologias:

- O *RandomForests*, um algoritmo de classificação, baseado em árvores de decisão, capaz de prever a partir de um modelo, criado por um processo de treino, para cada PTD, considerando apenas as variáveis existentes nesse PTD e o resultado para ele pretendido;
- O cálculo das menores distâncias euclideanas (método comparativamente mais rápido) entre o vetor das variáveis temporais existentes em cada PTD com *missing values* e o vetor das variáveis correspondentes nos centróides.

Os resultados apresentados na Tabela 8 mostram o número de PTD's alocados a cada *cluster* juntamente com os PTD's que não possuíam dados no período considerado (90 PTD's).

Tabela 8
Resultados da Classificação*:* número de PTD's alocados a cada *cluster* segundo a técnica da Distância Euclideana e segundo o algoritmo de classificação RandomForests.

Cluster	**Distância Euclideana**	**RandomForests**
Sem Cluster	90	90
3	6403	6153
2	1317	2118
1	2632	2081
Total	10442	10442

Após aplicação dos dois métodos, optou-se por comparar qual a percentagem de PTD's que tinham sido alocados ao mesmo *cluster* com o RandomForests e o cálculo da menor distância euclidiana e observou-se que aproximadamente 88% dos PTD's (9188) foram colocados no mesmo *cluster* independentemente do método utilizado. No entanto, foram assumidos os resultados da Classificação obtida a partir do Random Forest por serem mais fidedignos e específicos para cada PTD, assumindo sempre que para PTD's com poucos dados, os resultados poderão não ser tão precisos quanto desejávamos. Correções na alocação destes PTD's poderão ser realizadas posteriormente com acesso a maior quantidade de dados desse PTD, utilizando o mesmo algoritmo. Em termos práticos, é aconselhável utilizar o método da distância euclidiana mínima já que é mais rápido e os resultados obtidos não diferem significativamente dos resultados obtidos com o RandomForests.

Como resultado final desta etapa, conseguiu-se atribuir todos os PTD's a um dos 3 *clusters* obtidos após o *Clustering* e, consequentemente, construir os perfis tipo correspondente ao período de Novembro de 2016 a Outubro de 2017.

7.5. Construção de Perfis

A última etapa deste trabalho consistiu em construir perfis típicos de consumo para todos os PTD's, de modo a possibilitar o preenchimento de falhas

a partir das séries temporais obtidas com os centróides. Depois de alocados todos os PTD's aos seus *clusters*, foram então definidos os perfis típicos para o período de 2016 e 2017.

Visto que estas previsões constituem um enorme volume de dados é apresentada apenas uma parte do *dataset* com as previsões (Tabela 19).

Tabela 9
Exemplo da tabela que contém as previsões para o período de Novembro de 2016 a Outubro de 2017

Cluster	Média do Consumo Mensal	PTD_ID	01-11-2016 00:00	01-11-2016 00:15
2	0,364109239	1823D2032000_TP1	82,99107326	82,99107326
1	0,133978629	0711D2003800_TP1	0,474460789	0,474460789
1	0,133978629	1110D1025500_TP1	13,68267282	13,68267282

Convém referir que na definição dos perfis para o ano seguinte, se tinha considerado obter a constante de proporcionalidade, dividindo não as médias, mas a soma das potências ou potência total de cada PTD_ID no período anterior e a soma das potências do cluster a que cada PTD_ID pertencia. No entanto, havendo PTD_ID's que não tinham o mesmo número de variáveis temporais (leituras) que o centróide do seu cluster, tal não foi possível. Assim sendo, optou-se por calcular antes uma constante de proporcionalidade obtida pelo quociente entre as potências médias por período de 15 minutos (soma das potências a dividir pelo número de leituras) e a média das coordenadas de potência do centróide.

7.6. Análise de Resultados

O acesso aos dados reais dos diagramas de carga de Novembro de 2016, possibilitou a análise do erro percentual absoluto associado à nossa previsão. Dado que nem sempre os PTD's que aparecem nos dados são os mesmos, a comparação foi realizada apenas entre PTD's cujos registos estavam presentes em Novembro de 2015 e em Novembro de 2016. Os erros foram calculados para cada PTD_ID. Foi obtido um histograma (Figura 13) para os PTD's em Novembro de 2016 cujo erro médio entre previsões e dados reais era de

aproximadamente 30,2%. No entanto, tendo em conta o primeiro histograma da Figura 13 que revela a existência de possíveis *outliers* acima dos 10%, realizou-se um novo histograma que tinha em consideração apenas os valores que se encontram a menos ou equivalente a dois desvios padrão da média (Figura 13). Neste segundo caso, o erro médio diminui para cerca de 27,3%.

Finalmente, pretendeu-se verificar se o erro diminuía ainda mais se se considerasses apenas as previsões relativas a PTD's completos, já que logicamente os PTD's com mais falhas de registos deveriam ser mais difíceis de prever. No entanto, verificou-se que os erros não diminuíram significativamente, nem mesmo quando os possíveis *outliers* foram excluídos. Nestes casos, o erro obtido com todos os dados foi de 30,2% e sem *outliers* foi de 27,2%.

Concluiu-se então que o erro associado ao cálculo de estimativas utilizando os perfis obtidos através do *Clustering* estaria entre os 27,3% e os 30,2%, resultados que consideramos positivos. Anteriormente, na Secção 7.1, verificou-se que existia um erro de 15% associado ao cálculo de estimativas de valores de potência em falta durante o período máximo de uma hora, utilizando o método de regressão linear. Obter um erro de apenas 30,8% (praticamente o dobro do erro associado à estimativa efetuada para falhas de 1 hora), associado à utilização dos perfis obtidos com o *Clustering* de 58050 PTD's com apenas 3 *clusters*, foi então considerado um resultado bastante positivo, admitindo-se então que este método poderá ser viável para a validação de dados.

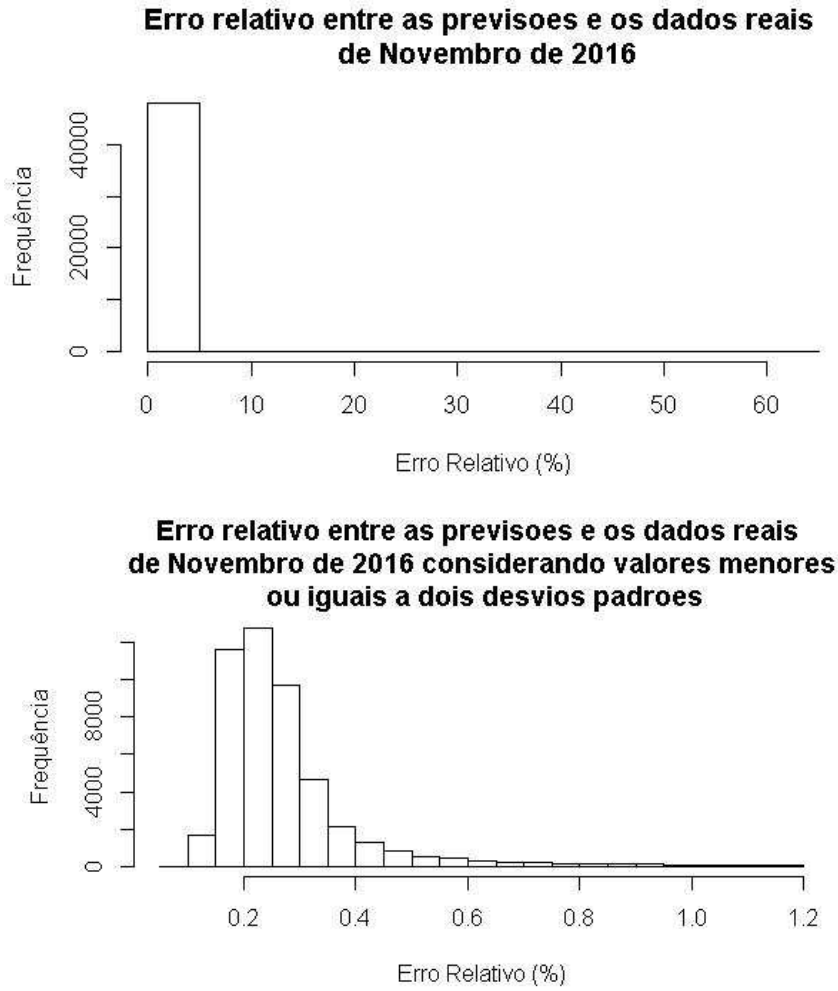


Figura 13

Erro Absoluto Percentual, apresentado sob a forma de um histograma, entre previsões e dados reais de Novembro de 2016. No primeiro histograma foram considerados todos os erros calculados enquanto no segundo histograma foram excluídos possíveis *outliers*, que se situassem além da média dos erros mais dois desvios padrões.

8. Conclusão

Este trabalho foi desenvolvido visando colmatar as falhas de registos dos valores de potência fornecidos pelos diagramas de carga dos PTD's e validar valores lidos.

Os resultados obtidos permitiram concluir que a utilização do *Clustering* para a segmentação de PTD's, e consequentemente a construção de perfis típicos representativos desses PTD's, permitia inferir os valores em falha para os diagramas de carga desses mesmos PTD's. Os resultados evidenciaram um erro médio entre dados reais e estimativas dos valores de potência em falta para Novembro de 2016 de aproximadamente 27,3%.

Apesar de vários obstáculos que foram surgindo, nomeadamente o grande volume de dados, o *hardware* disponível e certas limitações dos softwares utilizados, como a utilização da memória RAM pela ferramenta R, considerou-se que os resultados obtidos foram positivos e que poderiam ser utilizados para inferir os valores de registo em falta dos diagramas de carga. Neste sentido, a construção de perfis representativos de todos os PTD's possibilita não só a validação de dados para estimar valores de potência em falta, mas também a deteção de valores anómalos e a deteção de mudanças de comportamento ou padrões nos PTD's, algo que se poderá traduzir em novas tendências de comportamento.

Futuramente, considera-se que será relevante a realização de novos cálculos do valor do erro médio entre dados reais e previsões obtidas para os meses posteriores a Novembro de 2016, para verificar que, de facto, o erro não se altera significativamente, podendo a análise de *Clustering* ser utilizada para efetuar previsões. Além disso, seria interessante classificar novos PTD's segundo os 3 perfis representativos obtidos e, ainda, avaliar detalhadamente se existem características próprias (geográficas, tendências de consumo, entre outras) de cada segmento obtido a partir do *Clustering* que permitam a caracterização de cada grupo de PTD's.

Bibliografia

(ESMIG), E. S. (2011). A guide to smart metering. Bruxelas.

Acito, F., & Khatri, V. (2014). Business Analytics: Why now and next? *Kelly School of Business*, 565-570.

Ahmad, S. (2011). Smart metering and home automation solutions for next decade. *Proceedings of emerging trends in networks and computer communications conference*, (pp. 22-24). Udaipur.

Albert, A., & Ragajopal, R. (2013). Smart Meter Driven Segmentation: What Your Consumption Says About You. *IEEE Transactions on Power Systems*, 4019-4030.

Albert, A., & Rajagopal, R. (2013). Smart Meter Driven Segmentation: What Your Consumption Says About You. *IEEE Transactions on Power Systems*, 4019-4030.

Albert, A., Rajagopal, R., & Sevlian, R. (2011). Poster Abstract Segmenting Consumers Using Smart Meter Data. Seattle, Wa, USA.

Alejandro, L., Blair, C., Bloodgood, L., Khan, M., Lawless, M., Meehan, D., . . . Tuji, K. (2014). Global Market for Smart Electricity Meters: Government Policie Driving Strong Growth. *Office of Industries Working Paper U.S. International Trade Commission*.

Anderson, D. R., Burnham, K., & Thompson, W. (2000). Null Hypothesis Testing: Problems, Prevalence and an Alternative. *J. Wildl. Management*, 912-924.

Chen, J., Li, W., Lau, A., Cao, J., & Wang, K. (2010). Automated Load Curve Data Cleansing in Power Systems. *IEEE Transactions on Smart Grid*, 1.

Davenport, T., & Patil, D. (2012). Data Scientist: the sexiest job of the 21st century. *Harvard Business Review*.

De Silva, D., Yu, X., Alahakoon, D., & Holmes, G. (2011). Incremental Pattern Characterization Learning and Forecasting for Electricity Consumption using Smart Meters. *IEEE*, 807-812.

Depururu, S. S., Wang, L., & Devabhaktuni, V. (2011). Smart meters for power grid: Challenges, issues, advantages and status. *Renewable and Sustainable Energy Reviews*, 2736-2742.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence.*, 37-54.

Flath, C., Nicolay, D., Conte, T., van Dinther, C., & Filipova-Neumann, L. (2012). Cluster Analysis of Smart Metering Data. *Business & Information Systems Engineering*, 31-39.

Forgy, E. (1965). Cluster Analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 768-780.

Gillon, K., Brynjolfsson, E., Mithas, S., Griffin, J., & Gupta, M. (2012). Business Analytics: Radical Shift or Incremental Change? Assoc. Information Systems. *Digital Innovation in the Service Economy*. Orlando, Florida, USA: AIS Electronic Library.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 1-24.

Group, E. S. (2011). A guide to smart metering. Brussels.

Guo, G. (2003). K-NN Model-Based Approach in Classification. 986-996.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 107–145.

Hashem, I. A., Yaqoob, I., Salimah Mokhtar, N. B., & Samee Ullah Khan, A. G. (2014). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 98-115.

Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A Unified Foundation for Business Analytics. *Decision Support Systems*, 130-141.

Jain, M., & Verma, C. (2014). Adapting k-means for Clustering in Big Data. *International Journal of Computer Applications*, 19-24.

Jovic, A., Brkic, K., & Bogunovic, N. (2014). An overview of free software tools for general data mining.

Kádár, P. (2011). Smart meter based energy management system. *International Conference on Renewable Energies and Power Quality (ICREPQ'11)* (pp. 1160-1163). Las Palmas de Gran Canaria (Spain): RE&PQJ.

Kaufman, L. a. (1987). Clustering by means of Medoids. In Statistical Data Analysis Based on the L1–Norm and Related Methods. *North-Holland*, 405-416.

Kim, Y.-I., Shin, J.-H., Song, J.-J., & Yang, I.-K. (2009). Customer Clustering and TDLP (Typical Daily Load Profile) Generations Using the Clustering Algorithm. Asia: IEEE T&D.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* Springer.

Kwac, J., Flora, J., & Rajagopal, R. (2014). Household Energy Consumption Segmentation Using Hourly Data. *IEEE Transactions on Smart Grid*, 420-430.

LaValle, S. L. (2011). Big Data, Analytics and the Path from Insights to Value. *MITSloanManagementReview*.

Liu, X., & Nielsen, P. S. (2015). Streamlining Smart Meter Data Analysis. *Proceedings of the 10th Conference on Sustainable Development of Energy, Water and Environment Systems.* International Centre for Sustainable Development of Energy, Water and Environment Systems.

McLoughlin, F., Duffy, A., & Conlon, M. (2015). A clustering approach to domestic electricity load profile. *Applied Energy*, 190-199.

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1-2.

Provost, T., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Mary Ann Liebert, INC*, Vol1 No.1.

Verma, M., Srivastava, M., Chack, N., Kumar Diswar, A., & Gupta, N. (2012). A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Engineering Research and Applications (IJERA)*, 1379-1384.

Wu, X., Kumar, V., Quilan, J., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge Information Systems*, 1-37.

Xiong, D. (2007). A Local Density Based Spatial Clustering Algorithm with Noise. *Information Systems* (pp. 978-986). Elsevier.

Ylijoki, O., & Porras, J. (2016). Conceptualizing Big Data: Analysis of Case Studies. *Intelligent Systems in Accounting, Finance and Management*, 295-310.

Zerhari, B., Lahcen, A. A., & Mouline, S. (2015). Big Data Clustering: Algorithms and Challenges. *Conference: International Conference on Big Data, Cloud and Applications.* Tetuan, Morocco.